A SYSTEMS DEVELOPMENT OF A PROTOTYPAL STATISTICAL MODEL FOR THE LONG-RANGE PREDICTION OF DAILY CLIMATOLOGICAL MEASURES FOR A SPECIFIC GEOGRAPHIC LOCALITY

A THESIS

Presented to

The Faculty of the Division of Graduate Studies

By

Edwin Anthony Wray

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Industrial Engineering

Georgia Institute of Technology

November, 1977

A SYSTEMS DEVELOPMENT OF A PROTOTYPAL STATISTICAL MODEL FOR THE LONG-RANGE PREDICTION OF DAILY CLIMATOLOGICAL MEASURES FOR A SPECIFIC GEOGRAPHIC LOCALITY

Approved:

Cecil G. Johnson, Chairman

Douglas C. Montgomery

Paul G. Mayer

Date approved by Chairman December 5, 1977

PREFACE

This thesis was prompted by Johnson's concept (18) that daily climatological measures can be effectively described by probability distributions. Probabilities of occurrence can be obtained from empirical data. These probabilities can be refined by using accurate data and suitable methods.

A preliminary study of Johnson's concept was undertaken by Garrison (15). The study involved the development of the probability distribution of daily average temperature. Garrison used data from the period January 5 through March 26 for the years 1945 to 1965. The results of the study indicate that the method is valid.

The primary reason this subject is important in industrial engineering is the influence of climatological variables on operating systems. Effective systems design requires an understanding of the operating environment and its variability. The models developed in this thesis will provide a probability of the climatological measures being less than or equal to any value. This is an improvement over empirical probabilities that can result in improved systems planning. Another reason this subject is important is to order these forecasts into a format so a less experienced practitioner can predict with as good or better results than the experienced person.

The same reasons make this subject important in other fields of engineering. Civil engineers need climatic information for water run-off in designs, pollution dispersion, and water collection and evaporation in reservoirs. Climatic probabilities are needed for long-range scheduling of construction activities. Finally, engineers need probabilities of sunlight, cloud cover, temperature, and other climatic factors for design of solar energy systems.

ACKNOWLEDGEMENTS

This subject was originally suggested by Professor Cecil G. Johnson. His help extended far beyond this idea. He gave comprehensive guidance and encouragement. He has helped make this research an excellent learning experience for me. For all this, I am truly grateful.

Dr. Douglas C. Montgomery provided insightful comments and suggestions about the statistical aspects of this thesis. His support contributed to a sense of confidence that the research would be successful.

Dr. Paul G. Mayer participated as a member of my reading committee. His expert knowledge from a different field of engineering extended the scope of this research and I appreciate his professional insights.

I thank Claudine Taylor, Joyce Williams, and Cindy Disharoon for their fine professional efforts in preparing the final manuscript. Their professionalism and patience provided a polish that enhances the readability of the report.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONT'D.)

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

INTRODUCTION

This thesis is concerned with the development and test of a statistical predictive model for climate. Climate is defined as the prevailing or typical weather conditions of a place over a period of years. Climate should not be confused with weather. Weather is defined as being the general condition of the atmosphere at a particular time and place.

The systems approach to climate involves the modeling of the major climatological measures. In this case, systems is used to encapsulate all relevant measures of climate at one location. This concentration upon a systems model is in contrast to traditional modeling of single measures such as precipitation, pressure, etc.

The major thrust of this thesis is the development of a statistically valid model of daily climate. With such a model, it will be possible to forecast the probabilities of daily climate for each day of the year. Use of this model can be made without concern or knowledge of short range forecasts. The probability of a daily climatological measure is the critical information needed for long range planning.

Such a model is not related to the typical long term weather forecasts. Long dry spells, wet spells, periods of lower temperature are of no concern ih this model. Fluctuations of that nature are short term weather changes and are reflected in historic data used to develop the model. Thus they are included in the probability of the daily weather measures achieving the forecasted values. Once this model has been validated, it can be used to statistically describe every day of the year with the assurance that the forecast of fifty years from the present is as likely to be accurate as next year's forecast.

Generally, statistical models can be made more valid by increasing the volume of accurate data. The addition of data from the same station location and instrumentation should improve the validity of the model. However, in this specific case, the location of the recording station has made a significant change in location and height. Thus new data from this recording station might introduce measurement errors. This effect must be considered before adding more data.

This study is a prototype for all the days of the year. To accomplish this task, it is necessary to choose the days used in this study so that the different seasonal weather patterns are included. These days should be selected in such a way that all conditions are studied with little unnecessary overlapping. Sixteen days are used in this study to represent the different seasonal conditions.

These 16 days were selected around the two equinoxes and the two solstices. The four approximate dates are related to the yearly movement of the sun with respect to the earth's equator. Each date represents the beginning of a season. These dates are approximate because the solar year is 365¼ days long. This causes the time and sometimes the date of occurence to change from year to year in a four year cycle. The winter solstice occurs December 21 and is the first day of winter. On this date the sun is at its farthest southern point and the span of daylight is the shortest. The vernal equinox occurs

March 21, and is the first day of spring. At this time the sun is moving north across the equator. The summer solstice occurs June 21 and is the first day of summer. On this date the sun is at its farthest northern point and this is the longest period of daylight. The autumnal equinox occurs September 23 and is the first day of autumn. The sun is then moving south across the equator. These four dates are related to the movement of the sun and the sun has a strong effect on weather. Therefore, these dates represent significant physical effects on weather.

The solstices and equinoxes represent periods of changing climate. Eight additional days around these change periods are modeled in order to cover periods which may be unstable or different from the remainder of the year. These days are chosen twenty days before and after the four major solar events. These dates are, with the respective change point: December 1, (December 21), January 10; March 1, (March 21), April 10; June 1, (June 21), July 11; and September 3, (September 23), October 10. Thus the dates of the seasonal change and dates both before and after are modeled.

The four seasonal mid-points are modeled. It is believed that using these dates, a third type of climate will be analyzed. This third type should be more stable and representative of each season. These dates are the approximate mid-points between the solstices and equinoxes. These are: for winter, February 4; for spring, May 6; for summer, August 7; and for autumn, November 6. With the addition of these dates, all 12 months are represented at least once.

If these dates have been properly chosen, all major variations in daily data will be included in the analysis. Using these data in the analysis and development of statistical models will isolate any seasonal differences that would pertain to modeling. These differences could require the use of more than one statistical model to predict a daily parameter for each day of the year. If such cases do exist, they have a very great bearing on the extension of the models for these 16 days to the other 349 days of the year.

Statistical methods are used in the development of the prototype models. These statistical methods permit the reduction of a large volume of data to a small number of statistical parameters. Statistical methods are utilized with the emphasis on describing the phenomena rather than the power of the method. The development of new statistical methods is not an element of this study.

Serial correlations between data are considered between years. These correlations have a bearing on the independence of the data used in developing a statistical distribution function. The identification of a trend or lack of trend is a part of this study. Both trends and serial correlations between the yearly data values have a bearing on the use of statistical models.

There are some statistical methods that are excluded from consideration. These methods are based on correlations which make them more suitable for short-range prediction. Bayesian methods, time-series analysis, and Markov-chain models are three methods that fall into this category. Modeling with statistical distributions is a reasonable first step. The use of statistical distributions precludes the use of the above methods because correlations between years must not exist. If the data do have significant correlations, the three correlation-based methods would be logical extensions of this thesis.

The models used in this study are predictive. Daily models, developed in detail, are a means to predict climatological measures. These models are developed from historical data having no discernible serial correlations or trends.

There are models used in meteorology that are not used in this study. Some of the meteorological models use heat transfer, solar energy, primitive equations, simulations, or statistical-dynamical relationships. The above examples deal with physical relationships and cause and effect. As such these models deal with relationships that are beyond the scope of this thesis.

The meteorological models listed above attempt to explain how weather happens. This is not possible to accomplish using historical climatological data. Therefore, no attempt is made in this study to explain how or why weather happens.

Short-range forecasts can be developed from existing weather conditions using meteorological models. Such models are capable of providing general weather predictions for future months. Local forecasts of several days or a week are based on developing weather patterns over the entire country. Short-range predictions are normally more specific than the long-range prediction of climate. Such short-range predictions are projections based on existing conditions and are thus beyond the scope of this thesis.

If this prototypal model is found to be valid and useful, it can be extended to include all the days of the year. This extension would be a refinement of the model.

Daily measures are modeled to provide more accuracy. By modeling each day rather than each week, month, season, or year, variations due to seasons are reduced. Seasonal variations appear as gradual changes in magnitude of the measures as the time of year changes. Larger time spans of months, seasons, or years include more of these changes. A large time span could have a considerable spread in the daily data values used to develop the average. Thus the one value would be an inaccurate predictor for each day of the period. Daily models provide more accurate predictive models.

Data for 22 daily climatological measures are readily available from a government document, Local Climatological Data, published monthly by the United States Department of Commerce. These 22 measures are good descriptors of climate. Meteorologists at the United States Weather Bureau developed and refined the measures of climate in response to the users' needs. Thus these measures have empirical validity.

Twelve of these measures are included in this prototype thesis. These 12 were chosen as the most significant descriptors of climate. Most significant means most useful in this case. There may be cases where some of the other measures would be useful. As a prototype, this model attempts to prove the feasibility of the concept for the 12 measures at one geographic locality. Additional measures can be modeled to provide a climatological model that is better tailored to the specific requirements or locality. However these 12 measures should be sufficient for most needs.

Twelve of the 22 measures listed in the source document are modeled. These climatological measures are defined as follows:

1. Precipitation is all liquid and frozen water collected during the period midnight to midnight. Precipitation is measured to the nearest one hundredth of an inch.
2. Average wind speed is the average of the hourly wind speed readings for the period midnight to midnight. This is measured in miles per hour to the nearest tenth.
3. Maximum wind speed is the fastest wind speed recorded in the period midnight to midnight. It is measured in whole miles per hour.
4. Prevailing wind direction is the average of the 24 hourly wind direction observations for the period midnight to midnight. This is reported as one of the sixteen points of the compass or to nearest 10 degrees.
5. Maximum temperature is the highest dry bulb air temperature recorded in the period midnight to midnight. It is reported in whole degrees, Fahrenheit.
6. Minimum temperature is the lowest dry bulb air temperature recorded in the period midnight to midnight. This is reported in whole degrees, Fahrenheit.
7. Average temperature is the median of the maximum and minimum temperatures for the period midnight to midnight. It is reported in whole degrees, Fahrenheit.
8. Barometric pressure is the station observation for barometric pressure taken at 1300 hours. It is measured in inches of mercury to the nearest hundredth.
9. Relative humidity is the station observation for relative humidity taken at 1300 hours. It is a whole percentage.

10. Cloud cover is the average of the hourly cloud cover observations taken during the period sunup to sunset. It is measured in tenths of sky coverage.
11. Thunderstorm or distant lightning is recorded when observed or heard. This is at least one occurrence on the day. It is reported by letter or number codes.
12. Daily sunshine is the amount of sunshine striking the recording station. This is measured in hours and minutes.

Ten climatological measures were not included in this study. The reasons for the exclusion were varied. One, sky conditions, required a complex observation code to describe the type of cloud cover. The type of cloud cover did not appear to be as useful as the tenths of cloud cover. Therefore the type of cloud cover was not included in the model. Ceiling and visibility were not included because their usefulness is restricted to airline operations. Wet bulb temperature was not included because it is an observed quantity used to calculate relative humidity and dew point. Dew point is a value calculated from wet and dry bulb temperatures. It is a direct measure of the water vapor pressure. Because the usefulness of dew point is small, it was not included in the model. The quantity of snow, ice, and sleet on the ground at 0700 hours is available for each day. Atlanta does not have much frozen precipitation and so this measure would normally be zero and of little use in the model. The direction of the fastest wind was not included in the model. This climatological measure normally represents the direction of a strong gust of wind and not a true prevailing direction. Sky cover for the entire day, midnight to midnight, includes the measure of sky cover during daylight. The sky cover during daylight is also available. Although the daily sky cover is important to celestial observations, the sky cover during daylight has a greater effect on human activity. Therefore the sky cover during the daylight period was included in the model to represent this measure of climate. Degree days were not included in the model. Degree days represent the difference between the average daily temperature and the reference temperature of 65°F. Primarily the degree day is used as a measure of heating and cooling. The average temperature is included in the model. Therefore the probability of exceeding or falling short of the reference temperature could be computed from the cumulative distribution for average temperature. To include degree days in the model would be a duplication of this other information. Although some of these 10 measures might be difficult to model, each could be included in an expanded climatological model if needed.

The geographic locality used as a source of data is the Atlanta airport. A United States Weather Bureau station has been located at the airport since November 1, 1928. It is located at approximately 33°39' north latitude, 84°26' west longitude and is approximately 1000 feet above sea level. The recording station has occupied three sites from its inception until June 1, 1976. The elevation has varied from 973 to 1010 feet above sea level. The distances moved between the first and second sites, to the present site are 500 feet south and 0.9 miles, west, southwest respectively. Thus the weather station has been collecing weather data for a relatively long period and has not moved long distances.

The Atlanta climate is important to this study. The climate is classified humid subtropical. The lowest recorded temperature is -9°F and the highest recorded temperature is 103°F. Average monthly precipitation varies from 2½ inches to nearly six inches. The average yearly rainfall is about 48 inches. Average yearly snowfall is 1½ inches with a record snowfall of 8.3 inches occurring January 23, 1940. Winds average about nine miles per hour. This is not an extreme climate.

CHAPTER II

LITERATURE SEARCH

Introduction

The findings of this literature search are presented for each of the 12 climatological measures. Articles on each climatological measure are discussed with respect to the time spans modeled, the statistical methods used, and any extremes of climate which might affect the usefulness of the results.

The discussion of the time spans is meant to provide a brief overview of the modeling efforts for each measure. Conclusions about the modeled time spans are presented prior to discussing the statistical distributions.

Methods used in modeling the various measures are presented in the section on statistical distributions. This section provides the detailed information for each statistical distribution that was used to model the data. Unexplained findings of the articles that were briefly stated in the discussion on time spans are supported and discussed in detail. This section also ends with a presentation of the conclusions drawn from the statistical distributions used.

Each section on a climatological measure is ended with a discussion of the climate of the locations used in the articles. This discussion is primarily concerned with pointing out the similarities or dissimilarities of the climates of the locations modeled and Atlanta, Georgia. Extreme differences in climates can affect the relative magnitudes or frequencies of data values. Such differences can affect the useability of distributions and thus affect the ability to apply results achieved in one location to another location.

The chapter is finished with a discussion of the systems approach to the modeling of climate. In this section findings on the presence or lack of presence of the systems approach in the literature is presented. It is hypothesized that this approach to modeling more than one measure for a location has not been used prior to this thesis.

Precipitation

Time Span

Yearly. Mooley and Rao (28) modeled annual rainfall in India. Both the normal and gamma distributions were tried. The gamma distribution was found to be superior to the normal distribution for this purpose.

Seasonal. Mooley and Rao (28) modeled seasonal rainfall in India. As in the case of annual rainfall, the gamma and normal distributions were compared. The gamma distribution fit the seasonal data at all stations. In contrast, the normal distribution did not fit the seasonal data at a significant number of stations.

Monthly. Mooley (29) used statistics to model monthly summer monsoon rainfall in Southeast Asia. The Pearsonian distributions were investigated for fit to the data. The distributions tested were the normal, Type II, gamma, Type VIII, Type IX, exponential, and Type XII. Of all the distributions used, the gamma distribution had the best fit.

Weekly. Friedman and Janes (14) used a modified gamma distribution to model one week of summer data. The modification consisted of excluding zero precipitation amounts from the estimation of the gamma distribution parameters. These zero values were used to develop a probability of no rainfall. Unfortunately the authors based the choice of distribution entirely on works by other authors, in another geographic area.

<u>Daily</u>. Skees and Shenton (36) modeled daily precipitation using the gamma distribution. This treatment involved grouping daily precipitation amounts without regard to the day of the year. The authors concluded that the large variation in daily rainfall amounts caused the gamma distribution to provide a poor fit.

Mielke (26) used both the gamma and kappa distributions to model daily rainfall. The daily values used in this study were not stratified by the day of the year. In addition, only non-zero rainfall amounts were used as data points. Both the gamma and kappa distributions were found to fit the data. The author made the point that the kappa was the better distribution for detecting small-scale changes in precipitation because of its closed form.

<u>Other Time Spans</u>. Barger and Thom (2) calculated the probability of drought occurring during a period of one to sixteen weeks. A drought was defined as being a rainfall deficit. The gamma distribution was used to model the n-week rainfall amounts. The gamma distribution proved to be a good fit.

<u>Conclusion</u>. Based on the studies found in the literature survey, daily precipitation data have not been modeled by day of the year. There is considerable evidence that such a technique is viable. Therefore the statistical modeling of daily precipitation amounts is not rejected.

Statistical Models

Normal Distribution. Mooley and Rao (28) investigated the use of the normal distribution to fit seasonal and annual rainfall over India. Data from 53 recording stations were used. These stations provided data records of between 60 and 100 years. The authors tested the goodness-of-fit by using symmetry, kurtosis, and the chi-square test ($\alpha = 0.05$). Seasonal rainfall did not fit the normal distribution for 34 of the 53 stations. Annual rainfall was not fitted for 30 of the 53 recording stations. The authors concluded that "the distributions of seasonal and annual rainfall are not Gaussian over the major portion of India."

Mooley (29) studied the feasibility of using the normal distribution to model monthly summer monsoon rainfall. The period covered the months of June, July, August, and September. Data were used from 39 recording stations spread over Asia. Each station had a record of over 50 years. Goodness-of-fit was determined using a chi-square test, with a five percent significance level, and tests of skewness and kurtosis. Of the 39 stations and four months modeled, only 13 stations were fitted by the normal for June, 11 for July, 5 for August, and 6 for September. The author concluded that "the test for normality clearly indicates that monthly summer monsoon rainfall over Southeast Asia is not Gaussian."

Mooley (29) investigated three normalizing transformations for monthly summer monsoon rainfall. These transformations were the simple square-root, cube-root, and logarithmic. Only those data sets for stations and months that proved to be non-normal were used. The transformed distributions were tested for normality in the same manner as indicated in the preceding paragraph. The square-root and cube-root transformations led to normalization in about 70 percent of the cases. The performance of the logarithmic was effective in only 30 percent of the cases. "These transformations, therefore, have limited utility from the viewpoint of the normalizations of the frequency function for monthly rainfall over Southeast Asia during the summer monsoon."

Gamma Distribution. Mooley and Rao (28) used the gamma distribution to model seasonal and annual rainfall over India. The data were obtained from 53 recording stations with records of 60 to 100 years long. Both the chi-square tests ($\alpha = 0.05$) and the Kolmogorov-Smirnov tests indicated a good fit. The authors found that the gamma distribution was a good fit for all the data sets used.

The gamma distribution was used by Mooley (29) to model monthly rainfall in the Southeast Asia summer monsoon. More than 50 years of data were used from 39 stations. A chi-square goodness-of-fit test ($\alpha = 0.05$), with four to nine degrees of freedom, was used to test for fit. "In no case was the chi-square statistic significant at the 5-percent level." These findings were reinforced by the Kolmogorov-Smirnov test and the variance ratio test. Based on these findings the author accepted monthly summer monsoon rainfall as being gamma distributed.

Barger and Thom (2) used a gamma distribution to determine drought probabilities in six Iowa counties. These distributions were for n-week rainfall volumes where $1 \leq n \leq 16$. This period was May 17 to September 5, which is the corn growing season. Rainfall data from each county recording station were approximately 50 years in length. Chi-square tests ($\alpha = 0.05$), with two to six degrees of freedom, were used to determine fit. The n-week periods tested were 1, 2, 3, 4, 6, 8, and 10 weeks long. Only the two week period was a poor fit because "the bimodal nature of this particular histogram makes it difficult to get a good fit with a simple curve." Having used the gamma distribution to model rainfall data, the authors stated that the mode rather than the mean was the proper estimate of the most frequent amount of rainfall. This was due to the positive skewness of the rainfall distributions.

Friedman and Janes (14) used a modified incomplete gamma distribution to model seven day rainfall. This modification involved removal of zero rainfall values from the data. The zero values were then used to develop a probability of zero rainfall. The remaining data values were used to develop the gamma distribution parameters. Thirty years of rainfall data were used from Storrs, Connecticut. The seven day period was June 25 to July 1. No goodness-of-fit tests were used to check the fit of the gamma distribution to the data. The authors assumed that the conclusions of Barger and Thom (2) applied to Storrs, Connecticut. This assumption may not be valid because of the differences in geography and methods.

The modified incomplete gamma distribution is an interesting treatment of rainfall data. There may be cases where this method could give improved accuracy. Such a case would be rainfall probability distributions for arid locations. Such locations would have a relatively high number of zero values compared to an area receiving more normal rainfall. Another case where such a modification might be useful would be rainfall distributions for small time spans of a day or less. These smaller time spans would tend to have more zero values than a week or month. Such a high incidence of zero values could require this modification of the gamma or even another distribution that is better suited to model the zero values. The negative exponential is one such distribution. In any case the characteristics of the data will have to indicate the appropriate distribution.

Mielke (26) used the gamma distribution to model daily rainfall. The author chose 30 non-zero, non-seeded daily rainfall amounts collected during a cloud seeding experiment. The experiment took place at Climax, Colorado, and was conducted from 1965 to 1970. A chi-square test with two degrees of freedom was used to establish fit. The calculated chi-square value (1.191) was less than the theoretical chi-square value (5.99) for a five percent significance level and two degrees of freedom. Therefore the gamma distribution fits the non-zero daily rainfall data. These results are clouded by the non-random selection of data values.

<u>Kappa distribution</u>. Mielke (26) compared the two parameter kappa distribution to the gamma distribution. The same non-zero daily data were used from the Climax, Colorado, cloud seeding experiment. The chi-square value of 0.150 was again less than the theoretical chi-square value of 5.99 for a five percent significance level and two degrees of freedom. The two parameter kappa distribution fits the data.

The author makes the point that the kappa distribution has a closed form. This makes it easier to calculate expected frequencies rather than look up values in the tables. This is in contrast to the gamma distribution which requires interpolation of table values to give the expected frequencies. The author concludes that the kappa distribution may describe "certain observed precipitation amounts better than the gamma distribution." In addition, the kappa distribution is described as being easier to test for scale changes in cloud seeding experiments.

Exponential distribution. Mooley (29) investigated the use of the exponential distribution for modeling monthly Asian, summer monsoon rainfall. This was part of a larger study of Pearsonian distributions and their applicability to modeling rainfall data. First years of data from 39 recording stations were used to develop the gamma distribution parameters. When the value of the gamma distribution scale parameter is equal to one, the gamma distribution decomposes to the exponential distribution. Two locations had June rainfall modeled by the gamma distribution with scale parameters very close to unity. Based on the scale parameter value, the gamma distributions reduced to exponential distributions.

Pearsonian distributions (Types I, II, VIII, IX, XII). Mooley (29) investigated the fit of Pearsonian models to Asian monthly summer monsoon rainfall. The normal, gamma, and exponential distributions were discussed in previous sections of the chapter. Thirty-nine recording stations were used, each having more than 50 years of data. The analysis dealt with the relationship between the variables of the Type I distribution. This distribution was given as follows:

$$f(y) = \frac{m_1^{m_1} m_2^{m_2} \Gamma(m_1+m_2+2)\left(1+\frac{X}{a_1}\right)^{m_1}\left(1-\frac{X}{a_2}\right)^{m_2}}{(a_1+a_2)(m_1+m_2)^{(m_1+m_2)}\Gamma(m_1+1)\Gamma(m_2+1)}$$

The different distributions of the Pearsonian system were described by the author as being special cases of the Type I. Each distribution is characterized by the relationships between the Type I variables-- a_1, a_2, m_1, m_2.

The Type I distribution was checked for fit to the rainfall data. A chi-square test (α = 0.05) was used to test for fit with between two and five degrees of freedom. "The fit was good for 26 stations in June, 31 stations in July, 24 stations in August, and 23 stations in September."

The Type II distribution is a special case of the Type I. The Type I distribution approximates the Type II when $m_1 = m_2 = m$ and $a_1 = a_2 = a$. The author found that the Type I reduced to a Type II for two stations in June, three stations in July, one station in August, and one station in September.

The author stated with respect to Type VIII and Type XII:

> There was no case of Type I where the parameters m_1 and m_2 suggested that Type VIII would give a good fit and only one case (i.e. September rainfall at Lahore) where the parameters m_1 and m_2 of Type I suggested that Type XII may be a good fit.

The Type IX distribution was tested for fit using the monthly rainfall data. Chi-square goodness-of-fit tests (α = 0.05), with from three to six degrees of freedom, were used to determine fit. The Type IX provided a suitable fit to the rainfall data in only 13 cases.

Conclusion. There are three prime candidate distributions for modeling precipitation data. These are the gamma, modified gamma used by Friedman and Janes (14), and the two-parameter kappa distribution. All of these distributions should be checked to determine if they fit the data. The statistical goodness-of-fit tests are the determining factor as to which distribution is suitable for modeling daily rainfall data.

Location

Two studies dealt with modeling Asian rainfall. Mooley (29) modeled monthly summer monsoon rainfall for 39 stations in Southeast Asia. The farthest northern and eastern station was Tokyo, Japan. The farthest western station was Ahmadabad, India. The southern most station was Singapore. Included in these 39 stations were island, coastal, and inland locations. Mooley and Rao (28) modeled seasonal and annual rainfall at 53 stations in India. The rainfall in Asia tends to be heavily seasonal. The rainfall varies considerably by location. Caution should be used in directly applying these results to Atlanta, Georgia, without proper testing.

The remaining studies used data from the continental United States. Barger and Thom (2) used summer data from six counties in Iowa. Friedman and Janes (14) used data from one summer week in Storrs, Connecticut. Mielke (26) used data from Climax, Colorado. The locations used in the first two studies can be considered to be non extreme. The Climax, Colorado, location is at a much higher altitude than Atlanta, Georgia. Climax is in the mountains less than 30 miles from the highest peak (14,431 feet) in the state. With the exception of Colorado, the results can be expected to apply to Atlanta. It is still necessary to use proper methodology to select the suitable distribution.

Average Wind Speed

Time Span

Universal. Luna and Church (22) used the lognormal distribution to develop a universal wind speed distribution. Hourly and half-hourly data from 151 widely scattered stations were used to develop the distribution. By using the standard deviation of the universal lognormal distribution and the mean speed of a site, a local wind speed distribution can be obtained.

Conclusion. Although the concept of a universal wind speed distribution includes daily wind speeds, a daily wind speed distribution can probably provide additional accuracy. Any distribution that can be applied to such widely diverse area can be refined by restricting it to one area. Such a focusing for greater accuracy is one basis of this thesis.

Statistical Distributions

Lognormal. Luna and Church (22) used the lognormal distribution to model wind speed. Data sets of hourly and half-hourly observations from 151 recording stations in the United States, Wake Island, Puerto Rico, and Europe were used. Record lengths of the data varied from 10 to 15 years depending on the site. The data were massed without respect to season, month, or day for each site. Graphical methods were used to establish the fit.

The authors found that the lognormal distributions for each side were similar in shape without regard to season and height. Displacement toward greater or lesser speeds was the major difference. One conclusion of the authors was that the lognormal distribution was generally supported by the data used. With the standard deviation (1.9) of the universal distribution and the local mean speed, the local wind speed distribution can be developed.

There were other distributions that were excluded from consideration. The normal distribution was excluded because it involved negative values whereas wind speed data were strictly positive. All other distributions were excluded because "they were considered to be generally unfamiliar and therefore not as useful." Thus the lognormal was the candidate distribution by default.

Conclusion. Results of the Luna and Church study are not conclusive. Exclusion of all distributions from consideration, except the normal and lognormal, is not an exhaustive treatment of the subject. In addition graphical methods rather than statistical tests were used to establish fit. The gamma distribution has been used extensively in modeling precipitation data which are also strictly positive. Both the gamma and lognormal should be investigated for modeling average wind speed. Appropriate goodness-of-fit tests are available and should be used.

Location

The article by Luna and Church used data from widely scattered sites. A "universal" distribution that would fit 151 such sites would be easier to use but less accurate than a distribution developed for each site. A daily wind speed distribution for Atlanta should provide increased accuracy.

Maximum Wind Speed

Time Spans

Yearly. Thom (37) tested the fit of the Fisher-Tippett Type II distribution for modeling maximum wind gusts. This study was published as a design aid for structural engineers. As such it lacked detailed information about the statistical goodness-of-fit and the volume of data used.

Wood and Bowman (43) developed yearly peak wind profile probabilities for Cape Kennedy. A Fisher-Tippett Type I distribution was used to model peak winds at seven heights from 10 to 152.4 meters. Distribution fit was determined graphically on probability paper. The same methods were applied to the other time spans investigated in this study.

Wood, Palmer, Johnson, and Tyson (44) modeled extreme ground winds at Cape Kennedy. Yearly and other time spans were modeled using the same data and methods. Only three years of data were used to develop peak wind distributions for seven heights from 3.0 to 152.4 meters. The Fisher-Tippett Type I distribution was chosen for modeling the data based on the previous work by Wood and Bowman (43).

Seasonal. Wood, Palmer, Johnson, and Tyson (44) developed distributions of seasonal peak winds. These seasons were groupings of months and were not of equal length. Winter consisted of four months beginning with December. April and May made up the spring. Summer was four months beginning with June. The last two months, October and November, were the fall season. Cumulative distributions were developed at seven heights for each season and for six time spans (1, 5, 10, 15, 30, 60-day) within the season.

Periods of Two to 180 Days. Wood and Bowman (43) developed peak wind distributions for eight exposure periods of between two and 180 days. These periods were 2, 5, 10, 15, 30, 60, 90, and 180-days long. Peak wind distributions at the seven heights (10 to 150 meters) were developed for each of these time spans. Distribution parameters were displayed in tables. The cumulative distribution curves for each time span and height were displayed on probability paper.

Wood, Palmer, Johnson, and Tyson (44) developed extreme value distributions for the peak wind in 5, 10, 15, 30, and 60-day periods. Winds at seven levels of 3.0 m to 152.4 m were modeled for each time span. These time spans were modeled for each season and the year. The parameter values of the Fisher-Tippett Type I distribution were displayed in tables by height and time span for each of the four seasons and annual.

Daily. Wood, Palmer, Johnson, and Tyson (44) modeled daily peak winds using the Fisher-Tippett Type I distribution. Data and methods were identical to those used for other time spans in this article. Distribution fit was determined using probability paper. The results for all levels were displayed both tabular and graphically.

Wood and Bowman (43) extended their study to daily peak winds. Fisher-Tippett Type I distributions were developed at each level and displayed in tables and graphs. No volume of data was given in the article.

Okulaja (31) developed a frequency distribution of daily wind gusts for Lagos, Nigeria. These daily values were not stratified by day of the year. Plots of the data on extreme-value probability paper indicated that the Fisher-Tippett Type I (Gumbel) distribution fit the data.

Hourly and Less. Wood and Bowman (43) developed a model of one hour peak winds at seven levels of 10 to 152.4 meters. Cumulative distribution parameters were listed in a table for each level. Seven plots on a graph allowed the user to read off the probability of the maximum wind speed being less than some value x.

Wood, Palmer, Johnson, and Tyson (44) developed cumulative distribution functions for two-hourly increments and ten-minute peak winds. Distribution parameter values for the Fisher-Tippett Type I distribution were calculated for 152.4 and 18.3 meters, for the months January through December, by two-hourly increments from midnight on. The same information was developed for the ten-minute peak winds measured at 152.4 meters.

Conclusion. All the articles dealing with maximum or peak wind speed use one of the Fisher-Tippett extreme value distributions. These distribution were fitted to many different time spans using graphical methods. The daily time span, stratified by the day of the year, was not modeled. Sufficient information has been found to indicate that the method is viable. It remains for the methods to be applied to daily maximum wind speed data.

Statistical Distributions

Fisher-Tippett Type I. Wood and Bowman used an extreme value distribution to model peak winds at Cape Kennedy, Florida. This distribution was the Fisher-Tippett Type I or Gumbel distribution. Eleven time spans were modeled at seven levels. The time spans were one hour, 1, 2, 5, 10, 15, 30, 60, 90, 180, and 365 days. Each of the seven levels, 10.0, 18.3, 30.5, 61.0, 91.4, 121.9, and 152.4 meters, was modeled for each time span. No information was given about the quantity of data used to develop the distributions. Distribution fit was positive in all cases. This fit was represented as a straight line when the data were graphed on probability paper.

This study included many time spans and altitudes. It is unfortunate that the author was not clear in his use of the data. There was no information about how many data were used. In addition there is no clear statement of how the data were grouped. It is important to this thesis that the type of grouping or stratification be known. Without this information the results obtained by the authors have to be in question.

The study by Wood, Palmer, Johnson, and Tyson (44) was an extension of the study by Wood and Bowman (43). Again the authors used data from different heights at Cape Kennedy. These heights were 3.0, 18.3, 30.5, 61.0, 91.4, 121.9, and 152.4 meters. Some of the time spans differed from the original study. The time spans were ten-minute, 12 2-hour groups, 1, 5, 10, 15, 30, and 60-day periods. Seasons and months were also modeled. Only three years of data were used so the authors used the two-sample Kolmogorov-Smirnov test to determine if a three year subset of a 12 year sample taken at 10 meters was representative of the whole. "At the 5% level, the hypothesis of identical populations cannot be rejected." Based on this result, the three year data sample was used to construct probability distributions of peak winds for all time spans and heights. Models of the peak winds for all time spans were found by graphical methods to fit the Fisher-Tippett Type I (Gumbel) distributions. The parameter values of the distributions were calculated directly from the data and displayed in tables.

The three years of data were grouped into four seasons and annually. Peak wind distributions for 1, 5, 10, and 15-day periods were subsets of the seasonal groupings and the composite annual group. Thirty and 60-day periods were studied for the annual period. In addition the 10-minute and hourly peak wind speed distributions were developed by 2 hour group for the period January to December. The major deficiency of these data was that there was not enough.

There is no statement in the study that indicates the data were stratified by the day of the year. Instead the data appears to have been lumped together within the time spans. Even if the data had been stratified by the day of the year and daily distributions developed, three years is an insufficient volume of data.

Okulaja (31) used the Fisher-Tippett Type I distribution to model daily maximum wind gusts for Lagos, Nigeria. Daily data from 1948-1962 were used to develop the statistical model. These 4033 daily data were not stratified by the day of the year. Extreme-value probability paper was used to determine if the distribution fit the data. Half-widths of the 68 percent confidence band were drawn on both sides of the straight line through the data points. According to the author "the fit is considered good if two-thirds of the observed points lie within the control curves." The Lagos data satisfied this condition and the author concluded that the Fisher-Tippett Type I distribution fit the data.

Fisher-Tippett Type II. Thom (37) developed a distribution for maximum wind gusts using data from Fort Wayne, Indiana. The author used the maximum daily wind speed for each year. Although the author did not state how many years of data were used, from a plot of data points it appears to be more than 35 years. Distribution fit was determined graphically.

The author discussed why the Type II is the distribution for modeling maximum wind speed. A Type II distribution is bounded on the lower end and unbounded on the upper end. In contrast, the Type I distribution is unbounded at both ends. Therefore the Type II is better suited to model maximum wind speeds because they are always positive.

Iterative estimators of the Fisher-Tippett Type II parameters were presented in the article. An initial estimate of one parameter is obtained from a plot on probability paper. With the initial estimate, the equations can be solved iteratively for both parameters. A good description of the probability paper was included.

Conclusion. All of the articles on maximum wind speed involved extreme value distributions. In all cases the distribution fit was determined by plotting the data on probability paper. An approximate straight line plot indicates that the distribution fits the data. This was used for both the Fisher-Tippett Types I and II. The Type I was the most often used in the articles but it is unbounded on both ends. A Type II distribution was used in one article because it has a statistical lower bound. If maximum daily wind speed values are very large, there would be no problem with either distribution. If the maximum daily wind speed data are relatively low speed, the choice of distribution could be very important. The Type I distribution would involve negative values of wind speed using such data. For this reason the Fisher-Tippett Type II distribution is the prime candidate distribution for modeling maximum daily wind speed.

Location

The only article dealing with an extreme location is Okulaja (31). Lagos, Nigeria, is less than 7° north of the equator. Atlanta, Georgia, is approximately 33° north. The closeness of Lagos to the equator has an effect on weather and climate that is not felt in the more moderate climate of Atlanta.

Prevailing Wind Direction

Time Span

Universal. Luna and Church (22) developed statistics on wind direction in conjunction with the "universal" wind speed distribution. Sixteen compass points with the respective percentages of occurrence of hourly wind direction were used to model the wind directions for Alburquerque, New Mexico.

Conclusion. Nothing was found that specifically dealt with modeling daily average wind directions. There is no information contained in the article that would prove the concept of modeling daily prevailing wind direction is not feasible.

Statistical Distributions

Percentages. Luna and Church (22) used percentages of occurrence to model a "universal" wind direction profile. A 10-year summary of hourly observations was the data source. The airport weather station in Albuquerque, New Mexico, was the recording station. These data were tabulated into percentages of occurrence for the 16 points of the compass. The 16 points were N, NNE, NE, ENE, E, ESE, SE, SSE, S, SSW, SW, WSW, W, WNW, NW, and NNW. Percentages for the 16 compass directions were displayed in tabular form.

Court (6) described different types of wind roses that can be used to display percentages of occurrence by compass direction. A wind rose is a polar chart using bars, lines, or closed figures that are drawn proportional to the frequency of wind from each direction. Some roses add additional graphical codes to the ends of the bars or lines to represent the average wind speed from that direction. The wind rose is only a means for presenting results rather than a method of modeling daily wind directions.

Conclusion. The method of modeling wind direction used by Luna and Church (22) is simple and distribution free. Presentation of the results in table form is more accurate than the graphical wind rose. No other methods involving univariate or nondistributional statistics were found in this search. Therefore it would be reasonable to apply the methods used by Luna and Church to modeling daily average wind direction.

Location

Probabilities developed by Luna and Church (22) are distribution free. As such the method can be easily applied to any area. The Methods of the study can be used to model daily average wind direction in Atlanta, Georgia.

Maximum Temperature

Time Span

Monthly. Thom and Thom (41) developed and demonstrated a test to determine if the calculated monthly average maximum temperature had changed from the period 1921-50. In using this test, the authors assumed normality. Using the test to determine if two means are equal, the authors found that the monthly average maximum temperature was different for the periods 1921-50 and 1951-60.

Conclusion. Nothing was found that dealt with modeling any time spans other than monthly. There is no evidence that daily maximum temperature cannot be modeled.

Statistical Distributions

Normal. The statistical test presented by Thom and Thom (41) was based on an assumption that monthly average maximum temperature was normally distributed. According to the authors, monthly average, average minimum, and average maximum temperature can be statistically compared using this test. The two means used in the example were calculated from the periods 1921-50 and 1951-60. Data from March for Boise, Idaho, were used for the test. The approximation relates α, β, d (the difference in means), and the degrees of freedom. By using the approximation, the authors were able to use the normal probability tables rather than the tables of the more complicated noncentral t-distribution. A two-tailed test was used with $\alpha = 0.20$. Because of the costs to the users of changing the average monthly temperature values, "we have assumed that it would be as bad to revise a normal when it did not need it as it would be not to revise one when it really did need it. Hence, we wish to make the probability of a type 1 error equal to the probability of a type 2 error with alternative $|d|$ equal to 2.1°F." To equate the probabilities of type one and two errors, the authors calculated $\alpha = \beta = 0.20$. Using this method, the authors found that the monthly average temperature was not the same for both periods tested.

Location

Boise, Idaho, is located at an altitude of 2,704 feet in the arid valley of the Snake River. A 9,000 foot peak is located in the mountain range located less than 30 miles to the east of Boise. This area is not comparable to the geography of Atlanta, Georgia.

Minimum Temperature

Time Span

Monthly. Thom and Thom (41) assumed normality when they developed a statistical test to determine if the calculated monthly average minimum temperature value had changed from the period of 1921-50. An example of the test of two means was given using monthly average maximum temperature.

Conclusion. No articles were found in the literature that dealt with modeling daily minimum temperature. The one article that was found dealt only with testing two means and did not concern itself to the question of the feasibility of modeling any other time span. There is no evidence that daily minimum temperature cannot be modeled successfully.

Statistical Distributions

Normal. The statistical test developed by Thom and Thom (41) was based on an assumption of normality. This assumption was not supported by any tests. A test of two means was developed by the authors based on an approximation of Student's t-distribution that used the tables of the normal distribution. An example of the test was given using the monthly average maximum temperature and is described in the section dealing with that parameter.

Conclusion. The normal is one candidate distribution. Its candidacy is based partially on the fact that temperature can be either positive or negative. In the case cited in the article, the central limit theorem would indicate that the monthly average should be very close to normally distributed.

Daily average minimum temperature may not follow the same distribution. Since the daily values are medians and not averages, other distributions may be better suited. Two such distributions are the Fisher-Tippett Types I and II which are used to model extreme values.

Location

One article was found in the literature that dealt with minimum temperature. It was directed toward the description of methods that can be used to test for differences in monthly average values between two time periods. These methods are based only on the use of the normal distribution and are not affected by climate.

Average Temperature

Time Span

Monthly. Thom (39) used the normal distribution to fit monthly average temperature. Normality was an assumption that was not supported by any statistical tests. The author developed standard deviations of monthly average temperature at stations in the United States.

Thom and Thom (41) used the normal distribution to model monthly average temperature. The use of the normal distribution was not supported by statistical goodness-of-fit tests. A statistical test was developed to test if the average monthly temperature had changed from the value calculated for the period 1921 to 1950. The authors did not show an example using monthly average temperature.

Hourly. Wisner (42) developed frequencies of unfavorable temperatures for St. Louis, Missouri. Frequencies of temperature, exceeding or falling below a temperature, are presented in table form for two degree increments. These frequencies were tabulated by hour of the day for each month of the year. By using the tables, the percentage of the time that a temperature exceeds a set temperature (April to October) or is below a set temperature (October to April) can be determined.

Conclusion. No articles have been found that involved modeling daily average temperature. Nothing has been found that implies that the method will not work. Therefore, the method of statistically predicting long range average daily temperature is not rejected.

Statistical Distributions

Normal Distribution. Thom (39) developed the standard deviation of average monthly temperature. This was done for 150 first order stations of the United States Weather Bureau. These data were for a 30 year period from 1921 to 1950. The author states that "it is well known that the average monthly temperature is approximately normaly [sic] distributed." No statistical or graphical tests were shown to support this statement. The remainder of the article was devoted to developing the standard deviations and displaying them on charts as lines of constant value. Values taken from the chart can be used with the local mean temperature to develop degree days.

This article is interesting because it implies a consensus in the scientific community that average monthly temperature is normally distributed. It is unfortunate that no tests or references are given in support. An additional development of this study deals with the variability of the standard deviation.

> Perhaps the most noticeable property of the standard deviation of monthly average temperature is its conservatism relative to the location and mean. The range in mean for January from Florida to Montana is more than 60°F, but the range of the standard deviation is only about 8°F.

Thom and Thom (41) assumed that monthly average temperature was normally distributed. A test was developed to determine if the monthly average temperature had changed from the value for the period 1921 to 1950. No example test was conducted using monthly average temperature. An example use of the test was given using monthly average maximum temperature.

Percentages. Wisner (42) used percentages to model unfavorable temperatures for St. Louis, Missouri. Occurrences, in percent, were developed for temperature ranges divided into two degree increments. Fifteen years of data from the period 1940 to 1964 were used to develop the frequencies. April through October used a temperature range of 80° to 112° F. The tables showed the percentage of the readings that the temperature was equal to or exceeded the temperature increment. The period October through April used a temperature range from 0° to 30°F. Percentages for this period represented the portion of the time that the readings were equal to or below the temperature increment. In addition to being tabulated by month, the percentages were listed by hour of the day. These values were displayed in 14 tables.

Conclusion. Wisner's study does not directly apply to this thesis. It is included because the method can be useful. Unfortunately, the presentation of the results for 365 days would be cumbersome. It would require fewer numbers to provide probabilities of occurrence if a suitable distribution could be found. Therefore, this method is a last resort to be used if a probability distribution cannot be found.

Two of the articles stated that the normal distribution fits monthly average temperature data. These statements were not supported by appropriate tests and thus the conclusions cannot be used without testing. Therefore, the normal distribution should be investigated for modeling monthly average temperature.

Location

All three of the studies on temperature used data from recording stations in the United States. Because these data were not subjected to any goodness-of-fit tests, it would be useless to discuss the applicability of the results to Atlanta, Georgia. The applicability of the assumption of normality to Atlanta, Georgia, will have to be determined from the data. If the assumption of normality is valid, this thesis will provide the verification for models having the same climate as Atlanta.

Barometric Pressure

Time Spans

Daily. Madden and Sudah (23) developed the average and standard deviation of daily barometric pressure. Statistics were developed for each day of the year for Zurich, Switzerland. Parameter estimators for the normal distribution were used without identification or any test for normality.

Seasonal and Yearly. Madden and Sudah (23) investigated the independence and stationarity of seasonal and yearly averages of barometric pressure. Yearly data were found to be independent. The authors found that "short-term (on the order of a few years) changes in the statistical properties or climate of the data are small."

Conclusion. Madden and Sudah have developed statistical models similar to the ones proposed in this thesis. This effort was only a portion of the study. Daily distribution parameters were only graphically displayed which makes them difficult to accurately read.

Statistical Distributions

Normal. Madden and Sudah (23) used the parameter estimators of the normal distribution to develop the mean and standard deviation of daily barometric pressure. Data for this study were 12 observations, recorded in milibars (mb), averaged for each day to give 17,897 consecutive daily values. These data were recorded at Zurich, Switzerland, from January 1, 1901, through December 31, 1949, with no missing data. No tests, either graphical or statistical, were used to determine if the data were normally distributed.

A graph of daily average pressure and the standard deviation shows a wave-like variation over the year. The authors ascribe this variation to the "annual march of the sun's declination." This variation is very noticeable using millibars of pressure. From the graph it can be seen that the mean pressure varies about nine millibars during the year. Standard deviation varies about 10 mb. This variation would be negligible using inches because one inch equals 33.86395 mb.

Independence and stationarity of yearly average pressure were established using sign run length tests. The number of runs of the yearly means and standard deviations above and below their respective medians was compared to those expected from independent data. The expected number of runs was 25 for 49 independent values. Twenty-four runs for the mean were within one standard deviation (3.58) of the expected value, so yearly independence cannot be rejected. Twenty-eight runs of the standard deviation fail to reject yearly independence. The authors concluded that "any dependence that might exist between realizations is small, and that the assumption of independence is reasonable."

Conclusion. The article by Madden and Sudah is inconclusive. Proper tests must be run to determine the best distribution to fit daily pressure. The normal distribution is one candidate but it can include negative pressure values which are physically impossible. Another candidate distribution is the gamma which can be restricted to strictly positive pressure values. Statistical tests for goodness-of-fit can be used to determine the most suitable distribution for modeling daily pressure.

Another test can be used to establish independence of yearly pressure. Analysis of correlations and autocorrelations can determine if serial correlation is present. These tests should be able to resolve any questions about daily pressure that are left by gaps in the study.

Location

Zurich, Switzerland, is not a typical location. Zurich is at a much higher altitude than Atlanta, Georgia, and would thus have a lower average barometric pressure. In addition the Alps would have an influence on the local weather that would not be present in the Atlanta area.

Relative Humidity

Time Span

Monthly. Yao (44) investigated the modeling of relative humidity using the beta distribution. Sixty monthly data sets were selected from 204 (19 stations times 12 months). This screening was done to remove data sets that were spatially correlated. Using a chi-square test ($\alpha = 0.05$), only three out of the 60 monthly data sets were found to be not beta distributed.

Fifteen Day. Yao (44) used data from Washington, D.C., to investigate the use of the beta distribution to model 15 day average relative humidity. Six data sets of 15 day average relative humidity were calculated from daily readings from the sixteenth to the thirtieth of the month. A chi-square test ($\alpha = 0.05$) failed to reject any of the 15 day averages as not being beta distributed.

Ten Day. Yao (44) used 10 day averages to test whether relative humidity was fit by the beta distribution. These 10 day averages were calculated using readings from the sixth to the fifteenth of the month in Washington, D.C. Eleven such data sets were tested using a chi-square test ($\alpha = 0.05$) and all were accepted as beta distributed.

Five Day. Yao (44) in the same article tested five day averages of relative humidity. The five day averages were calculated using the first five days of the month in Washington, D.C. Seven data sets were checked using a chi-square test ($\alpha = 0.05$) and all were accepted as beta distributed.

Daily. The distribution of daily relative humidity was investigated by Yao (44). Observations were used from the thirty-first day of the month in Washington, D.C. Six daily data sets were tested. The chi-square test ($\alpha = 0.05$) rejected one data set as not being beta distributed.

Conclusion. The work by Yao (44) provides good evidence that daily relative humidity can be statistically modeled. Using data from only the thirty-first day of the month does not prove that the technique is applicable to the entire year. This thesis can provide more complete proof about the feasibility of using the method for the entire year.

Statistical Distributions

Beta. Yao (44) studied the use of the beta distribution for modeling different time spans of relative humidity. Monthly, five day, ten day, fifteen day, and daily time spans were investigated. All data sets were tested for fit using a chi-square test ($\alpha = 0.05$).

Monthly relative humidity was investigated for 19 stations over the continental United States. Sixty years of data were used. These data were from the 0700 and 1900 EST observations for each station. All data sets were stratified by location and month. The author screened the 204 (19 stations times 12 months) for spatial correlations using a critical correlation coefficient of 0.20. This screening reduced the number of data sets to 60. Only three of these 60 sets were rejected by the goodness-of-fit test. Analysis of the monthly distributions for Phoenix, Arizona, show that the p and q values of the beta distribution increase from March to a maximum in July. This indicates that the cumulative mean monthly relative humidity curves change with season at that location.

The remaining time spans, daily, five day, ten day, and fifteen day were modeled using data from Washington, D.C. Only the 0700 EST observations were used in developing the averages. Fifty-three years of data were used in this segment of the study. Six daily, seven five-day, eleven ten-day, and six fifteen-day data sets were developed and tested from the Washington data. Only one daily data set, January 31, failed to fit the beta distribution.

Conclusion. The study published by Yao (44) used good methodology in investigating the use of the beta distribution to model relative humidity. It seems highly probable that the beta distribution can be used to model daily relative humidity for Atlanta, Georgia. Proper tests will still have to be conducted to ensure that the results do in fact apply.

Location

Washington, D.C., is located near the junction of the Potomac and Anacostia Rivers. These rivers have an effect on humidity during the warmer months. Atlanta, Georgia, does not have the same proximity to large bodies of water. This difference may be a factor in using the beta distribution to model relative humidity in Atlanta.

Cloud Cover

Time Spans

Monthly. Falls (13) investigated the underlying theoretical statistical distribution of monthly world cloud cover. Ten of the 29 cloud cover regions of the world, including six covering the United States, were included in this study. The author found that the beta distribution fit monthly cloud cover data for all regions.

Hourly. Falls (13) used hourly data in his study of world cloud cover. Data for the hours 0400, 1000, and 1600 were modeled using the beta distribution. Using a Kolmogorov-Smirnov test, the author determined that the beta distribution fit hourly cloud cover data.

Daily. Thom (40) in a brief technical note, predicted the mean percentage of sky cover for March 7, 1970. This was presented as an aid for selecting a viewing site for a solar eclipse on that date. No information was given on how the results were arrived at.

Conclusion. The note by Thom (40) does not provide information on modeling daily cloud cover. In contrast, Falls (13) presents evidence that the technique is viable. It remains for this thesis to develop a daily average cloud cover model.

Statistical Distributions

Beta. Falls (13) used the beta distribution to model monthly and hourly cloud cover. Ten of the 29 world cloud cover regions were included in this study. Region 19 which includes the southeastern United States was one of these regions. Data for the study were obtained from three of the references used for the article. Assumptions from the three references about the area included in an observation were referred to by the author. "The assumption was made (1,2) [sic] that the weather observer can accurately record cloud-cover observations up to 27.80 km (15 n. mi) i.e., the cloud cover diameter is 55.60 km (30 mi.) with the station at the center." Observations of cloud cover are described as the amount or tenths of cloud cover. These cloud cover data were stratified by each month and the hours 0400, 1000, and 1600. The months were January, April, July, and October and they were described as "representative of the annual changes in the cloud distributions in the earth's atmosphere." The data, in tenths of cloud cover, were then grouped into five categories for analysis. Category one included zero tenths; two included one, two, and three tenths; three included four and five tenths; four included six, seven, eight, and nine tenths; and category five included only ten tenths. This was done to simplify the analysis. A Kolmogorov-Smirnov test was used to determine fit. Because there was no information about the sample size of the data, the author assumed n = 100 for all distributions. In his conclusions, the author states that the beta and normal distributions were tested for fit. "This investigation has shown that the underlying distribution of cloud cover is definitely not normal and, furthermore, the normal distribution is unbounded at both ends, which violates the constraints of our random variable." In contrast, the beta distribution was rejected in only seven of 160 cases at a five percent significance level. The author concludes that the beta distribution is the underlying model for world cover.

Conclusion. This study by Falls (13) strongly indicates that the beta distribution is the underlying model for cloud cover. Unfortunately the data used by the author came from other sources and is open to some question. Inability to use the actual sample size of the data in the calculations must have an effect on accuracy. This accuracy can be improved by using data properly collected from a good source. It is reasonable to assume that the beta distribution will fit daily cloud cover data and that this will be validated by testing.

Location

The study by Falls (13) used data from cloud cover regions of the world. Each region is considered to have representative homogeneous cloud cover distributions. It is reasonable to have inconsistencies in cloud cover in a region covering the entire southeastern United States. Inconsistencies could occur near large bodies of water or mountainous area. Such inconsistencies would add variability to regional statistics. Development of a cloud cover distribution for a specific locality should improve the prediction.

Thunderstorms or Distant Lightning

Time Spans

Monthly. Sakamoto (33) tested the fit of the Poisson and negative binomial distributions to monthly thunderstorm days at five stations in Nevada. Monthly thunderstorm days for May through October were fitted by the negative binomial distribution. November through April thunderstorm days were fitted by the Poisson distribution. The applicability of the distributions depends on the relative frequency of occurrence.

Falls, Williford, and Carter (12) investigated probability distributions for modeling "thunderstorm events" and "thunderstorm hits" for all months. A "thunderstorm hit" was defined as a thunderstorm passing over Cape Kennedy. The "thunderstorm event" was defined as a thunderstorm occurring at Cape Kennedy and surroundings. The authors concluded that the modified negative binomial and the negative binomial distributions gave a good fit to the thunderstorm hits and thunderstorm events respectively.

Yearly. Sakamoto (33) used the Nevada data to investigate the fit of the Poisson and negative binomial distributions to annual thunderstorm days. Annual thunderstorm days were found to follow the negative binomial distribution at all five sites. The probability distribution of annual thunderstorm days provided probabilities of having X number of thunderstorms in a year.

Conclusion. None of these studies deal with modeling the probability of one or more thunderstorms occurring on a given day. On the other hand, the modeling of daily thunderstorm probability was not proved impossible. It remains for this thesis to develop such a model.

Statistical Distributions

Negative Binomial. Sakamoto (33) successfully used the negative binomial distribution to model thunderstorms for annual and six high frequency months. Data were collected at five stations in Nevada from 1942 to 1971. Both the chi-square and Kolmogorov-Smirnov tests established distribution fit to the data.

Falls, Williford, and Carter (12) found the negative binomial distribution provided the best fit of thunderstorm events at Cape Kennedy. The data were collected at Cape Kennedy, Florida, for the period January 1957 through December 1967, a period of 11 years. A chi-square goodness-of-fit test was used to establish fit in all cases "where data were sufficiently large to admit its use." The negative binomial was found to be superior to the binomial and Poisson distributions for modeling thunderstorm events.

Poisson. Sakamoto (33) determined that the Poisson distribution was a good fit to data for six months of data at five sites in Nevada. These six months had a low frequency of thunderstorm occurrence. Data were collected from 1942 to 1971. Suitability of fit was determined using the chi-square and Kolmogorov-Smirnov tests.

The Poisson was found to provide a poorer fit to both thunderstorm events and hits in a study by Falls, Williford, and Carter (12). In the article the authors did not say that the Poisson did not fit the data. It was stated that the negative binomial and modified negative binomial distributions had the smaller computed chi-square values.

Conclusion. The articles found in the literature and described above, do not directly relate to modeling daily thunderstorms. When modeling daily thunderstorms, the value of the information lies with the probability of at least one on that date. The above articles deal with multiple thunderstorms occurring during the time span. It will be necessary to use methods which can describe the probability of at least one occurrence on the date. One candidate method is a simple percentage of occurrence.

Location

The two articles found in the literature use thunderstorm data collected in the United States. Five sites in Nevada and Cape Kennedy, Florida, were the recording stations used. There is nothing unusual about the climate of these sources that would be extreme.

Sunshine

Time Spans

Monthly and Yearly. Angell and Korshover (1) investigated long-term trends in percent sunshine. One hundred and three stations in the contiguous United States were divided into six regions. Mean-monthly values of percent sunshine and the respective standard deviations were used in the analysis. The authors determined that for the period 1950 to 1972, autumn percent sunshine decreased eight percent and spring percent sunshine increased three percent over the entire country.

Conclusion. The existence of a trend in the data can affect the feasibility of modeling with statistical distributions. If that is the case with daily sunshine duration, then this measure would have to be omitted from this model. Because of the results of the study by Angell and Korshover (1), special care is required in checking for trends or serial correlation in daily data from Atlanta, Georgia.

Statistical Distributions

Angell and Korshover (1) did not use a distribution to model daily sunshine. Instead they used the mean-monthly percent sunshine and the standard deviation. Twenty-three years of data from the period 1950 - 1972 were used from 103 stations having unbroken data records. These stations were chosen from 160 available because there had been no station relocations during the data period. The majority of these stations were located at airports near city outskirts. The stations were grouped into six geographic regions which were the Northwest, Southwest, North Central, South Central, Southeast, and Northeast. According to a map of the station locations, Atlanta, Macon, and Savannah were three of the 20 stations included in the Southeast region. The data period of 1950 to 1972 was chosen to minimize any biases due to the instrument changes made in the early 1950's.

The authors discussed the photoelectric sunshine switch which has been used to measure sunshine duration since the early 1950's. This instrument replaced the Marvin sunshine switch which was used from 1897 until the advent of the new equipment. The amplifier of the photoelectric sunshine switch (sunshine detector) has been made more sensitive in 1957, 1963, 1965, and 1972. Each change in sensitivity resulted in the detector turning on earlier in the day and turning off later in the day. Because of the limitations in the detectors, the observers had to correct for late turnon and early turnoff under clear sky conditions. Dust, pollutants, and smoke have an effect on the detector accuracy when the sun is near the horizon.

The data taken by the sunshine detectors were used to develop two statistics. A monthly mean value of percent sunshine (S) was developed for each month of the year. A standard deviation of the monthly averages was calculated for each month. No formula was given for the standard deviation. These statistics were used to analyze general characteristics of yearly sunshine duration, spatial correlations, and long-term trends.

Percentage sunshine duration was found to be cyclical. A maximum for the United States occurs in summer and a minimum in winter with the difference being about 20 percent. The annual variation is not symmetric and can vary between regions. Normally the autumn decrease in percent sunshine is more abrupt than the increase during spring. The Southeast is anomalous in that the maximum occurs during the late spring rather than during the summer. According to a graph of standard deviations by region and time of the year, the Southeast has a two percent variation between the low in May/June and the high in September/October. The maximum variation of standard deviation for any region was eight percentage points.

Spatial correlations were calculated for stations separated by 600 and 200 km. At 600 km apart the "correlations average about 0.5, although varying from 0.4 in the Southeast and Southwest to 0.6 in the Northeast. The relatively low values in Southeast and Southwest reflect the different S regimes at coastal and inland stations." At a distance of 200 km apart, 85 percent of the correlations were between 0.7 and 0.9.

The authors analyzed the data for long-term trends.

> The long-term trend has been examined by applying a 1-2-1 smoothing (2-1 and 1-2 at beginning and end of record,

respectively) to 3-year block average values of S (4-year averages at beginning and end of record). This procedure leads to a conservative estimate of the long-term trend.

All regions exhibited an autumn decrease in percent sunshine, whereas in the spring they exhibited an increase. These trends for the whole country are a three percent increase in spring and an eight percent decrease during the autumn. According to a table, 95 percent of the stations in the Southeast region showed a seven percent decrease in autumn percent sunshine between 1950 and 1972. For the year, there has been a 1.3 percent decrease in percent sunshine for the whole country from 1964 to 1972.

Using a 12-month running mean, the authors found that the bulk of the variation in data was "quasi-biennial." This variation in percentage of available sunshine in the contiguous United States showed an out-of-phase relationship with the 50 mb zonal wind at Balboa, Canal Zone. The correlation and 95 percent confidence limits between the wind at Balboa and percent of possible sunshine for the Southeast was -0.19 ± 0.13. The smallest calculated correlation was exhibited by the Northeast, -0.09 ± 0.12, the largest by the Northwest, -0.22 ± 0.18, and the United States Average was -0.21 ± 0.06.

Three explanations were given for the changes in percent sunshine.

1. There may be a long-term climatic trend associated with an expansion or shift in the location of the north polar vortex.
2. Aircraft contrails may have caused an increase in fairly thick cirrus cloudiness.

3. Pollution increases may have been large enough to affect the turning on and off of the sunshine recorder.

Conclusion

Although this study does not deal with modeling sunshine duration, it does present relationships between sunshine data values that have an effect on the modeling. The existence of trends in regional data does not necessarily mean that the trends are significant at all locations. Data for Atlanta, Georgia, will have to be analyzed to see if this trend is significant and would affect the use of statistical models.

In the absence of any studies dealing with statistical modeling of sunshine, it is reasonable to discuss some of the possibilities. Available daily sunshine is a deterministic value that has been computed for each day of the year. Modeling of the hours and minutes of sunshine for each day would be difficult and cumbersome. It would be easier to model the percent of available daily sunshine. The beta distribution has been used to model percentage of occurrence for other climatological measures. It would be a reasonable candidate for modeling percent of available daily sunshine. A model of percent of available and a table of the available sunshine by day, together would give probabilities of having less than x hours of sunshine.

Location

The one article found in the literature on sunshine did not deal with modeling. Results and conclusions obtained through the analysis dealt with regions of the country. Because of their size, the different regions could not be homogeneous in their topography. Therefore there would be variation, especially when the region included both coastal and inland recording stations. Thus the results for a region may not hold for all locations within the region. Data from Atlanta, Georgia, will have to be analyzed by itself to determine if there is a trend or serial correlation.

Systems Approach

Two articles were found in the literature that involved more than one climatological measure. Luna and Church (22) modeled average wind speed and wind direction. Thom and Thom (41) investigated minimum, average, and maximum temperature in the same study. This investigation involved statistical testing rather than modeling. In both cases the climatological measures studied were closely related.

No articles were found that modeled unrelated climatological measures in such a way as to describe the climate of one geographic locality. The only works that were found to exist on modeling total climate were developed by government agencies and consisted of high, low, and average values of climatological measures. Extreme and average values constitute an incomplete model because no information is presented about the variability of the historical data. Therefore it is concluded that the systems approach to modeling climate has not been used in conjunction with statistical models.

CHAPTER III

ANALYSIS AND DESIGN OF THE MODEL

Identical methodology was applied in the analysis of all but two of the climatological measures. Wind direction and thunderstorms only required the simple calculation of probabilities. Each step in analysis of the other 10 measures was built upon the results of the previous step. All data points were first graphed to allow a visual check for correlations. Next, two or more of the data sets, those with indications of correlation were analyzed mathematically for correlation. If these test data sets showed no correlation, frequency histograms were developed to facilitate the choice of candidate distributions. Parameters were then estimated for each of the statistical distributions under consideration as models. Goodness-of-fit tests were conducted using these parameters. If the goodness-of-fit tests showed that the theoretical distribution fit the data, the models were then validated. Each of these steps is described in more detail in the remainder of this introduction.

Data were collected from government publications supplemental by station logs (24). The primary source was Local Climatological Data (21) for the Hartsfield Airport Station in Atlanta, Georgia. This source contained daily climatological data from 1950 to the present. Because of changes in the reporting formats of relative humidity and barometric pressure data, station logs from 1965 to

1975 were used as a supplementary source. While using these station logs some minor errors in the published data were found and corrected. Station logs were the only source of data from 1940 to 1949.

Data sets of 33 and 35 years were used for this study. Thirty-five years of data were avilable for all measures except relative humidity and barometric pressure. Both of these measures were described by the 1300 hour observation. The observations were not begun until 1942. Thus only 33 years of data were available for relative humidity and barometric pressure. All data sets were complete with no missing data points.

The data sets were analyzed to determine if there were correlations in the data that would preclude the use of statistical distributions. First, the daily data sets for each measure were graphed. Using these graphs, two days for each measure were chosen for further analysis. These two days showed the indications of correlations between data. If these data sets did not prove to have serial correlations or trends, then the remaining 14 days for that measure could also be accepted as uncorrelated.

Each of the selected days was mathematically analyzed for trends and serial correlations. This was done by calculating (5) the auto-correlation and partial autocorrelation functions for each data set. The joint behavior of the correlation functions for a data set can indicate serial correlations or trends. If such correlations were found in the two days representing a climatological measure, the remaining 14 days could be analyzed in the same way. Because relatively small data sets can give inconclusive results using this method, a majority of the 16 data sets must show similar structure before the climatological measure can be excluded from further analysis.

Frequency histograms were developed for those climatological measures which did not have strong correlations. This step was excluded if it was obvious that statistical distributions would not be used to model a particular measure. Frequency histograms provided a view of the general shape of the frequency curve inherent in the raw data. Using these frequency histograms, candidate statistical distributions could be hypothesized to fit the data sets for each measure. Normally the number of distributions were restricted to one or two of the most likely candidates.

Distribution parameters were estimated for each statistical distribution which could fit the 16 daily data sets for each measure. The importance of good parameter estimates cannot be overstressed. A goodness-of-fit test uses the parameters to compare the theoretical statistical distribution to the data set and either indicates no fit or failure to reject. If the parameters are inaccurate the comparison will not be valid. The easiest parameter estimation methods were used for each statistical distribution. If the subsequent goodness-of-fit tests indicate a bad or borderline fit, then the parameters were re-estimated using maximum likelihood. Even when the best estimation formulas are used, insufficient data can cause inaccuracies. This may be a factor in this study because no more than 35 years of good data were available for most of the climatological measures.

Two goodness-of-fit tests were used to determine if a theoretical distribution fit the data. A Kolmogorov-Smirnov goodness-of-fit test was the primary method. This was chosen because of the ease with which it can be programmed and its independence from the number of groups. All distributions were tested with the Kolmogorov-Smirnov test and all conclusive results were accepted. The problem with the Kolmogorov-Smirnov test that necessitated the use of the Chi-Square goodness-of-fit test was the unavailability of critical values for all the statistical distributions used in this study. All of the distributions required the estimation of parameters which made the most widely published tables unusable.

Crutcher (7) provides tables of critical values for the gamma, normal, and Type I extreme value distributions. These values are tabulated for N=25, 30, and greater than 30, and alphas of .20, .15, .10, .05, and .01. Using the critical values detailed in this table, it is possible to set a very conservative critical value for distributions not included in the table. Critical values for N=35 and α=.05 vary from .150 for the Type I extreme value and gamma (Shape parameter greater than eight), and the normal distributions to a high of .164 for the gamma (Shape parameter equal to two). A very conservative critical value would then be .099. Therefore a Kolmogorov-Smirnov test for the beta, lognormal, or Type II extreme value distributions could be accepted so long as the critical value did not exceed 0.099. Values in excess of this, in the order of .250 and above, would be rejected. Test values between .100 and .250 would be checked using the chi square test. Thus obvious results could be accepted and questionable results mathematically checked at a known significance level. Borderline tests using parameters estimated by other than maximum likelihood would be rerun after the parameters had been reestimated. Given a case where two goodness-of-fit tests had been run using parameters estimated by different formulas, the best or most optimistic results of each test run would be used. An example might be where method of moments had provided parameter estimates that were better than maximum likelihood for five of sixteen cases as indicated by the goodness-of-fit tests. The results of the five successful moment estimated parameters would be used with the 11 maximum likelihood parameters to represent the 16 data sets for that climatological measure.

The models of the climatological measures were validated using data from 1975. This was only necessary for those statistical distributions which can model a measure. Validation consists of inputing the 1975 data into the statistical models and determining the probabilities of attaining the values. If a model of daily climate was not able to accommodate these data, then the model was not valid and must be rejected.

Precipitation

Precipitation data were graphed for the 35 years from 1940 to 1974. These graphs were characterized by a significant number of zero values. In every case the data sets consisted of more than one-half zero values. Day 4 had 19 zero values out of 35 points. Day 12 had 33 of 35 points equal to zero.

Days 1 and 8 were tested for trend and serial correlation. Day 1 proved to be negative. The second day tested, day 8, showed a weak correlation which may have been caused by small sample size. In either case the results were not significant and thus it could not be proved that correlation was present in the precipitation data. Thus, there were no correlations that would exclude precipitation from this study.

Distribution parameters have to be estimated from the available data. The presence of so many zero values makes it very difficult to estimate parameters and determine the suitability of distributions to model precipitation. Such characteristics of the data, greather than or equal to zero, coupled with the two most likely distributions, gamma and lognormal, produced a problem. The gamma and lognormal do not include any zero values. The probability of zero precipitation would require an addition to a lognormal or gamma distribution for rainfall. The remaining non-zero precipitation data were not enough to estimate distribution parameters.

It was decided to present whatever information available in these limited data in the most straightforward manner.

Probabilities of measurable precipitation are present in Table 1. For the purpose of this study, trace amounts of precipitation were recorded as zero.

Table 1. Probability of Measurable Precipitation

Day	Non-zero data	Probability	Day	Non-zero data	Probability
1.	12	.343	9.	15	.429
2	11	.314	10.	11	.314
3.	14	.400	11.	7	.200
4.	17	.486	12.	2	.097
5.	11	.314	13.	6	.171
6.	7	.200	14.	7	.200
7.	9	.257	15.	4	.114
8.	12	.343	16.	11	.314

Records for 1975 indicate that measurable amounts of rain fell on days 1, 2, 3, 5, 8, 9, 12, and 15. Of these eight days, the first six have a tabulated probability of precipitation in excess of .300. Days 12 and 15 have probabilities of .097 and .114 respectively. Although these two probabilities are small, the models do accept the 1975 data and therefore are valid.

Average Wind Speed

Thirty-five years of average daily wind speed data were graphed for each of the 16 days chosen. Days 1, 3, 6, 7, 8, 11, 12, 13, and 15 showed were basically stable (Figure 1). The other days showed different degrees of variability.

Days 2 and 9 were judged to have the highest probability of correlation. This hypothesis was not proved by analysis of the correlation function.

Histograms were developed from the data sets (Figure 2.). The shape of the histograms varied from left and right skews to symetric. Because the data are all greater than zero, two theoretical distributions could fit the data. The two distributions were the gamma and lognormal.

Parameters for both the gamma and lognormal distributions were estimated using maximum likelihood estimators (Appendix A-2, A-3). All data points were non-zero and thus contributed to parameter estimates. These parameter estimates and the results of the Kolmogorov-Smirnov goodness-of-fit tests are shown in Table 3.

Test results for the gamma distribution were evaluated using Crutcher's critical values. The results for the lognormal were evaluated using the conservative critical value and the chi-square goodness-of-fit test. Both the gamma and lognormal distributions did not fail to fit any of the 16 daily data sets.

The gamma and lognormal distributions were validated using the data from 1975. Probabilities for the gamma distribution varied from .108 for day 13 to .960 for day 12. Although both of the distributions fit the data, both do not give the same probabilities for the same data.

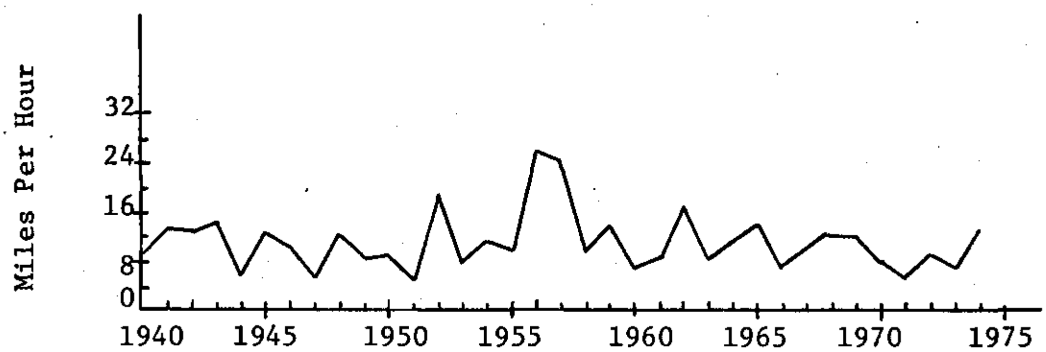

Figure 1. Average Wind Speed (Day 1)

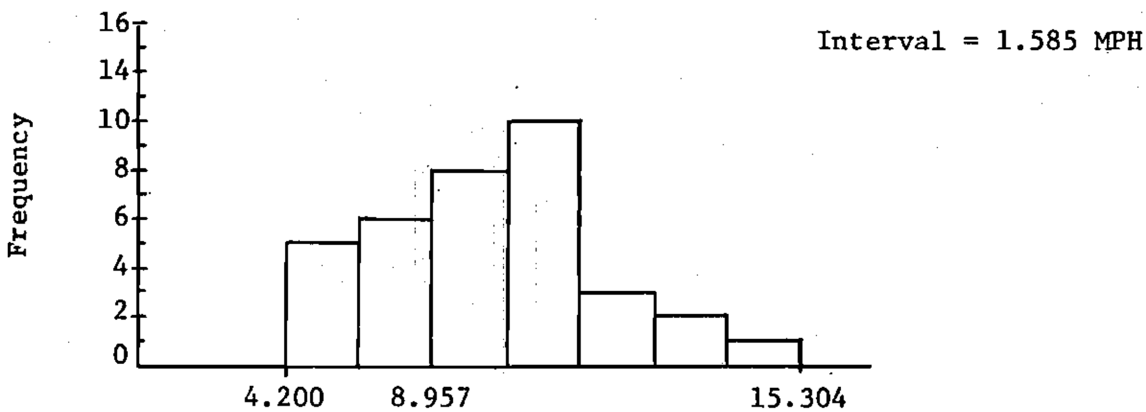

Figure 2. Average Wind Speed (Day 6)

Table 2. Average Wind Speed Models

	Gamma Distribution			Lognormal Distribution		
Day	Gamma	Beta	Test Statistic**	Mean	Variance	Test Statistic***
1	4.53	2.52	.118	2.358	.146	.087
2	4.70	2.37	.086	2.340	.142	.122
3	4.90	2.34	.099	2.371	.154	.079
4	4.54	2.60	.126	2.394	.162	.129
5	5.79	1.67	.129	2.211	.103	.159
6	6.68	1.28	.117	2.098	.094	.078
7	5.07	1.63	.091	2.046	.134	.085
8	6.07	1.28	.120	1.997	.099	.101
9	6.44	1.14	.086	1.948	.102	.060
10	5.84	1.18	.080	1.873	.114	.054
11	4.80	1.50	.117	1.908	.141	.119
12	5.97	1.42	.099	2.086	.102	.076
13	5.05	1.58	.070	2.009	.129	.090
14	4.18	2.13	.089	2.102	.170	.074
15	3.94	2.20	.111	2.071	.189	.084
16	4.92	1.98	.114	2.208	.125	.077

* Fails to fit

** If the value of gamma was greater than four, the critical value for gamma greater than or equal to eight was used. This value was the most conservative of all critical values available for the gamma distribution with N=35 and α=.05. Critical value = .150.

*** All Kolmogorov-Smirnov test statistics less than one were accepted. The remainder were tested using a chi-square test with α=.05.

Maximum Wind Speed

Graphs of the maximum wind speed data all showed variability. The days 1, 2, 3, 5, 6, 7, 11, 12, 13, 14, 15, and 16 all showed relatively small variations (Figure 3). The other four days showed larger variability. Nine days showed possible correlations or trends.

Days 1 and 11 were chosen as the data sets exhibiting the strongest indications of correlations. Neither of these two days produced correlation functions that indicated structure. Thus it was concluded that there was nothing that would preclude modeling these data.

Frequency histograms were developed for the 16 data sets. The normal, gamma, and Type II extreme value distributions were distributions that could fit these histograms (Figure 4). A normal distribution could possibly be used to model maximum wind speed if the speeds were large and thus the probability of zero or negative values of x were equal to zero. This was not the case with these data and thus the normal was excluded. The two remaining distributions under consideration were the gamma and Type II extreme value.

The parameters for both the gamma and Type II distributions were calculated using maximum likelihood (Appendix A-3, A-6). The parameters for both distributions are displayed in Table 3. These are the parameter estimates that were used in the goodness-of-fit tests.

The Kolmogorov-Smirnov tests results are displayed in Table 3. A critical value of .150 was used to test the gamma distribution for fit.

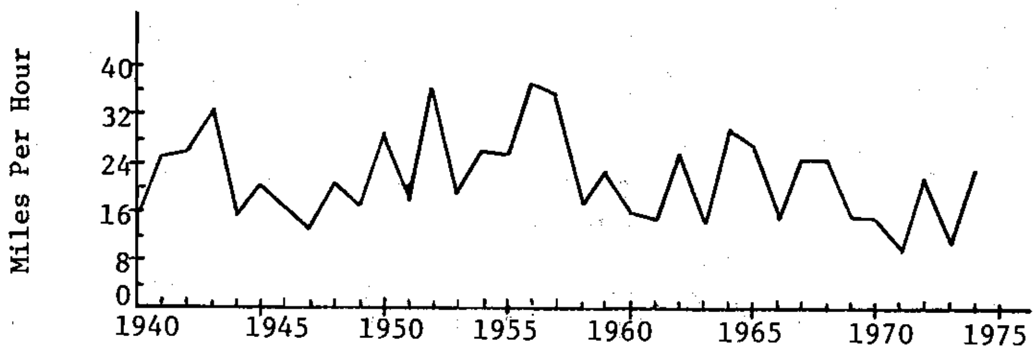

Figure 3. Maximum Wind Speed (Day 1)

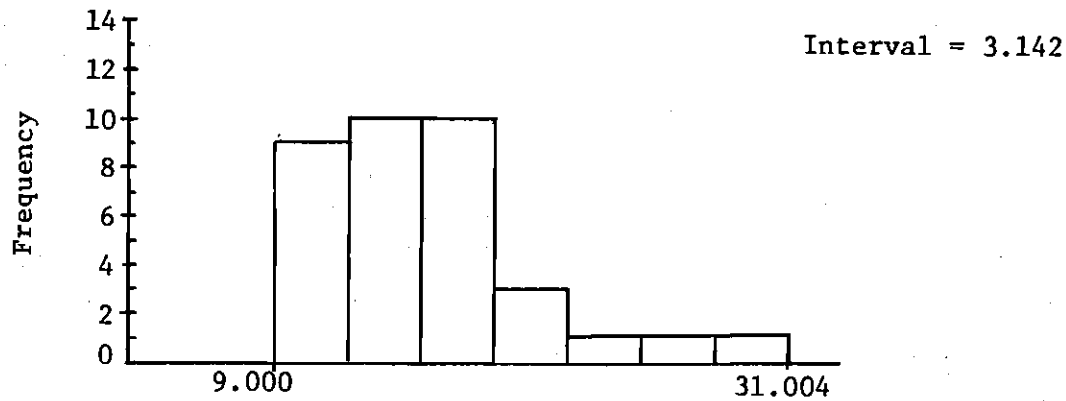

Figure 4. Maximum Wind Speed (Day 9)

This was at a significance level of .05. Only one day, day 12, failed the test. Likewise, day 11 failed to be fit by the type II extreme value distribution. Five days required chi-square tests to determine fit or no fit. One failure out of 16 cases cannot be considered be significant. Therefore both of these distributions were considered to be good models for daily maximum wind speed.

The models using the gamma and Type II extreme value distributions were validated. In this validation, the two cases where the distributions failed to fit the data, were excluded. The cumulative probabilities for the gamma distribution varied from .08 for day 13 to .982 for day 1. Cumulative probabilities for the Type II extreme value distribution varied from .003 for day 13 to .989 for day 12. The cumulative probabilities from both of these models are reasonable and thus the models are validated.

Table 3. Maximum Wind Speed Models

	Gamma Distribution			Type II Extreme Value Distribution		
Day	Gamma	Beta	Test Statistic**	Kappa	Theta	Test Statistic***
1.	5.54	3.82	.095	2.90	16.77	.096
2.	6.55	3.07	.111	3.49	16.50	.071
3.	5.41	4.00	.058	2.90	17.09	.094
4.	4.72	4.93	.068	2.73	17.89	.091
5.	6.24	3.03	.122	3.49	15.48	.077
6.	5.68	3.31	.094	2.73	14.87	.150
7.	6.12	2.91	.135	3.37	14.57	.093
8.	3.98	5.18	.086	2.68	15.36	.085
9.	7.72	2.00	.119	4.06	13.05	.110
10.	7.72	2.10	.115	3.53	13.51	.101
11.	6.15	2.39	.127	2.46	11.66	.174*
12.	5.88	2.62	.181*	3.74	12.58	.084
13.	6.51	2.37	.134	3.75	12.75	.086
14.	4.99	3.27	.090	2.60	12.66	.102
15.	4.76	3.60	.105	3.37	13.53	.082
16.	6.75	2.57	.097	3.89	14.36	.085

* Fails to fit

** If the value of gamma was greater than four, the critical value for gamma greater than or equal to eight was used. This critical value was the most conservative of all critical values available for the gamma distribution with N=35 and α=.05
Critical Value = .150

*** All Kolmogorov-Smirnov test statistics less than one were accepted. The remainder were tested using a chi-square test with α=.05.

Wind Direction

Wind direction data represent the portion of the compass from which a wind is blowing. As such these data do not lend themselves to graphing by years and tests for serial correlations. Histograms, estimation of parameters, and goodness-of-fit tests required for most of the other measures are not required for the tabulation of probabilities by compass point.

The sixteen points of the compass and their associated degrees are presented in Table 4. Tabulation of these data was made more difficult by changes in the reporting of wind direction. During the 35 years in which these data were taken, the presentation of wind direction has changed from compass points to degrees. These later data are only reported to the nearest 10 degrees.

Converting the degrees to one of the 16 compass points introduced a bias. Because 360 degrees did not divide evenly by 16, it was necessary to increase four points by 10 degrees each. Thus northwest, northeast, southeast, and southwest span 30 degrees each, while the other 12 points span 20 degrees. Bias of some kind could not be avoided because of the inexactness of the wind direction reporting systems.

The probabilities of wind direction were calculated using 35 data points. These are reported to four decimal places to reduce rounding errors. All probabilities were calculated for each day and presented in Table 5.

The average wind directions for the 16 days were gathered for 1975. The occurrence of these directions was compared to the probability of occurrence. All days but day 11 had directions that have a probability of occurrence of .0286 or better. Day 11 had an east, southeast wind which had zero probability according to the tabulated probabilities. This result was not surprising considering the relatively small number of data used in calculating the probabilities. It is reasonable for all zero probabilities to disappear from the table as the number of samples increase. Thus, this model is accepted as being valid for predicting wind direction.

Table 4. The 16 Points of the Compass

Direction	$<$	$\leq$
North	350	10
North Northeast	10	30
Northeast	30	60
East Northeast	60	80
East	80	100
East Southeast	100	120
Southeast	120	150
South Southeast	150	170
South	170	190
South Southwest	190	210
Southwest	210	240
West Southwest	240	260
West	260	280
West Northwest	280	300
Northwest	300	330
North Northwest	330	350

Table 5. Wind Direction Probabilities

DAY	N	NNE	NE	ENE	E	ESE	SE	SSE	S	SSW	SW	WSW	W	WNW	NW	NNW
1	.0286	.0000	.0286	.0000	.0571	.0857	.0286	.0286	.1143	.0000	.0857	.0571	.0286	.0000	.3429	.1143
2	.0000	.0000	.0000	.1714	.1143	.0286	.0000	.0000	.0857	.0000	.1143	.0286	.0571	.1143	.2571	.0286
3	.0000	.0000	.0000	.0000	.1714	.0571	.0857	.0000	.1143	.0000	.0571	.0571	.0286	.0571	.2571	.1143
4	.0571	.0000	.1143	.0000	.1714	.0000	.0000	.0000	.0857	.0286	.0857	.0000	.1429	.1714	.1143	.0286
5	.0000	.0286	.0286	.0571	.1429	.0857	.0571	.0571	.0857	.0286	.0571	.0000	.0286	.0571	.2571	.0286
6	.0286	.0000	.0000	.0286	.0857	.0571	.0857	.0286	.0571	.0000	.1429	.0857	.1429	.0571	.1429	.0571
7	.0571	.0000	.0857	.0857	.0286	.0000	.0571	.0286	.0857	.0571	.2286	.0286	.0286	.0286	.1429	.0571
8	.0286	.0000	.0000	.0286	.0857	.0000	.0857	.0000	.1143	.0571	.1143	.0286	.1143	.0571	.2571	.0286
9	.0286	.0000	.1143	.1143	.0286	.0000	.0857	.0286	.1143	.0000	.0571	.0286	.1429	.0857	.1714	.0000
10	.0000	.0000	.0000	.0571	.0857	.1143	.0857	.0857	.0286	.0286	.0571	.0571	.0857	.0571	.2000	.0571
11	.0571	.0000	.0857	.0286	.1714	.1143	.1429	.0000	.0857	.0286	.0286	.0286	.0286	.0571	.0857	.0571
12	.0286	.0000	.1714	.2286	.1143	.0286	.1143	.0000	.0286	.0000	.0000	.0000	.0286	.0286	.1714	.0571
13	.0571	.0000	.0857	.0286	.1714	.0286	.0857	.0286	.0571	.0000	.0286	.0000	.0286	.0857	.2286	.0857
14	.0857	.0000	.0286	.0571	.0286	.1143	.0000	.0000	.0571	.0286	.0571	.0000	.0571	.0286	.3714	.0857
15	.0857	.0000	.0286	.0571	.1714	.0286	.0571	.0000	.0000	.0286	.0286	.0571	.1429	.0857	.2000	.0286
16	.0286	.0000	.0000	.1143	.0857	.0571	.0571	.0286	.0286	.0000	.1429	.0286	.0571	.1143	.2571	.0000

Maximum Temperature

Maximum temperature data sets were graphed chronologically. These graphs were fairly stable for the days 1, 5, 7, 8, 9, and 10 (Figure 5). The remaining 10 days showed different degrees of variability.

Days 1 and 3 were chosen for correlation tests. Both data sets were inputed into the computer program and the correlation functions calculated. Neither of the days showed significant correlations. Therefore the maximum temperature data are assumed to be uncorrelated.

Having accepted the data sets for modeling, frequency histograms were developed. The histogram for day 10 (Figure 6) is a good example of the 16 histograms. Considering the range of the data and the shapes of the histograms, the normal and Type I extreme value distributions were chosen for further analysis. Both theoretical distribution can accept negative, zero, and positive data, and have curves similar to the general shapes of the histograms.

Parameter estimates for both distributions were developed from the data using maximum likelihood (Appendix A-1, A-5). Estimates for both distributions are presented in Table 6. These values were used in all subsequent goodness-of-fit tests.

Kolmogorov-Smirnov critical values were available for both the normal and Type I extreme value distributions. The critical value for N=35 and five per cent significance level was .150 for both distributions. Results of the goodness-of-fit test are displayed in Table 6. The extreme value distribution fit 12 of the 16 days. All four of the

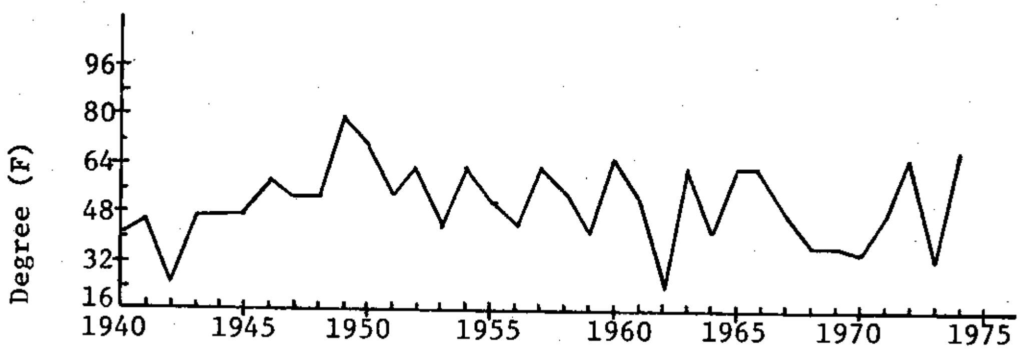

Figure 5. Maximum Temperature (Day 1)

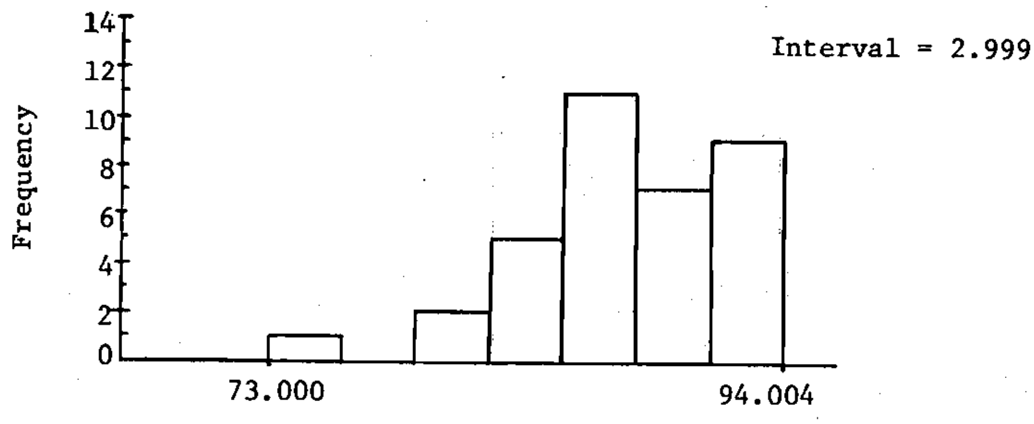

Figure 6. Maximum Temperature (Day 10)

Table 6. Maximum Temperature Models

	Normal Distribution			Type I Extreme Value Distribution		
Day	Mean	Variance	Test Statistic**	Theta	Psi	Test Statistic**
1.	49.9	163.20	.076	12.30	43.53	.081
2.	51.6	88.70	.097	8.16	47.11	.063
3.	58.9	89.93	.080	9.01	54.09	.084
4.	61.7	100.50	.099	8.96	56.77	.063
5.	69.9	64.92	.067	8.53	65.66	.137
6.	78.7	55.22	.091	7.23	74.96	.105
7.	84.3	41.96	.057	7.17	80.89	.102
8.	87.8	25.25	.119	6.60	85.01	.189*
9.	87.1	16.98	.066	5.22	84.90	.169*
10.	87.1	21.30	.080	5.39	84.67	.162*
11.	85.9	42.96	.078	7.95	82.32	.184*
12.	80.5	38.31	.099	6.35	77.25	.108
13.	75.5	38.94	.086	7.27	72.17	.142
14.	64.7	47.48	.066	7.82	60.98	.144
15.	55.5	77.56	.088	7.84	51.17	.087
16.	53.6	88.35	.070	9.45	48.79	.101

* Fails to fit

** Critical value = .150 (N=35, σ = .05) failures were in sequence and all in the summer months (day 8 is June 21 and day 11 is September 3). These results would make the normal distribution the preferred model at this time. The extreme value model required further analysis prior to any acceptance.

The model using the normal distribution was validated. Cumulative probabilities varied from a low of .039 for day 2 to .937 for day 4. Both of these values together represent a range of reasonable probabilities. Therefore the model using the normal distribution was accepted as a valid predictor of maximum temperature.

Minimum Temperature

Graphs of minimum temperature data sets showed considerable variability. With the exception of days 1, 2, 7, 8, and 12, the graphed data fluctuated considerably (Figure 7). In addition, 13 of the days gave at least a minor indication of correlation.

Days 3 and 8 were judged to have the greatest chance of correlation. The correlation functions of both days did not show any significant correlation. Therefore, it was concluded that daily minimum temperature data were not correlated and could be modeled using statistical distributions.

Using the frequency histograms of the 16 data sets, two candidate distributions were selected. Both the normal and Type I extreme value distributions can match the histograms and cover the range of data. Both of these distributions were analyzed as possible models of minimum temperature.

Two methods were used to estimate the parameters of the two distributions. Maximum likelihood methods were used to estimate the normal distribution parameters (Appendix A-1). Estimators developed by Gumbel (16) were used to estimate the parameters of the Type I extreme value distribution (Appendix A-1). This was necessary because no maximum likelihood estimators were available for cases of minimum value. These parameter estimates were the most accurate using available methods and are shown in Table 7.

Critical values for the Kolmogorov-Smirnov goodness-of-fit test were available for both distributions. All 16 data sets were fitted

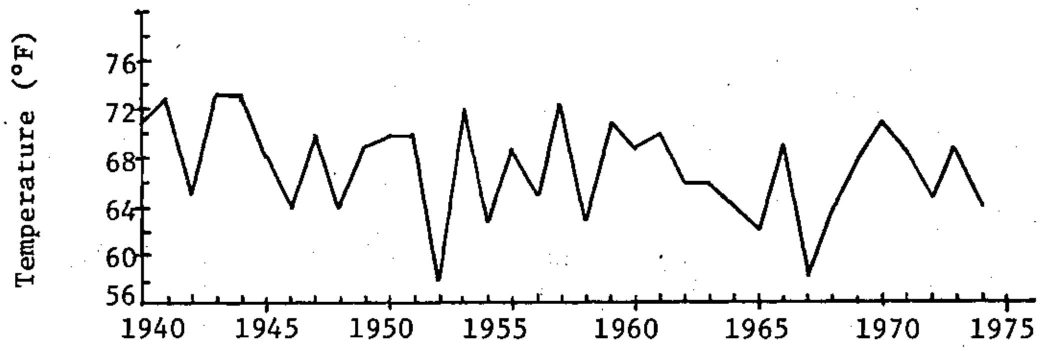

Figure 7. Minimum Temperature (Day 11)

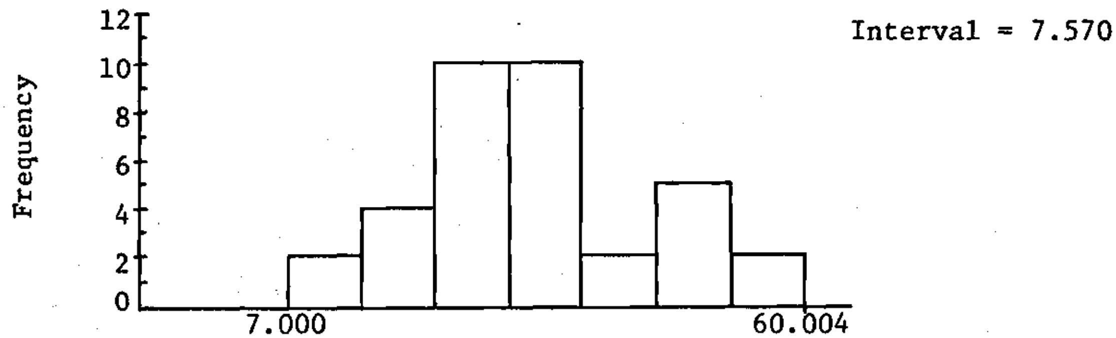

Figure 8. Minimum Temperature (Day 1)

by the normal distribtuion. Four days were not fitted by the type I extreme value distribution (Table 7). These were days 15, 16, and 1 which include the months of December and January. The fourth day was day 8 which is June 21. The test statistic of .155 for day 8 barely exceeded the critical value of .150. Failure of three data sets in a row and in cold months raises doubts about the overall suitability of the type I extreme value distribution for modeling minimum temperature. These doubts may be eliminated by the addition of more data. At present the normal distribution appears to be a suitable model for predicting daily minimum temperature.

Model validation was conducted for only the normal distribution. The type I extreme value distribution was not successful in modeling some of the colder days. The cumulative probabilities calculated using the normal ranged from .177 to .956. The capability of the model to include the validation data and develop reasonable probabilities establishes the validity of the model.

Table 7. Minimum Temperature Models

	Normal Distribution			Type I Extreme Value Distribution		
Day	Mean	Variance	Test Statistic**	Theta	Psi	Test Statistic**
1.	32.5	143.70	.132	9.34	37.59	.183*
2.	32.4	98.35	.086	7.73	36.58	.129
3.	38.6	75.15	.082	6.76	42.28	.123
4.	41.8	66.41	.073	6.35	45.20	.129
5.	48.3	62.03	.096	6.14	51.60	.097
6.	55.7	46.82	.093	5.34	58.63	.086
7.	62.9	32.42	.056	4.44	65.31	.066
8.	67.1	11.19	.108	2.61	68.52	.155*
9.	68.4	12.36	.095	2.74	69.91	.119
10.	69.3	10.31	.096	2.50	70.61	.054
11.	67.2	16.33	.078	3.15	68.90	.119
12.	61.8	32.89	.082	4.47	64.24	.081
13.	54.0	46.63	.073	5.32	56.88	.103
14.	42.4	55.45	.097	5.81	45.57	.111
15.	33.8	53.13	.124	5.68	36.87	.167*
16.	35.6	107.10	.132	8.07	39.93	.185*

* Fails to fit

** Critical value = .150 (N=35, α=.05)

Average Temperature

Thirty-five daily average temperature values were plotted by year. Half of the days plotted displayed strong variability (Figure 9). The other days, 2, 7, 9, 11, 13, 14, 15, and 16 showed less variability.

Days 5 and 7 were chosen for the correlation test. The structure of the correlation function for day 7 was inconclusive. The correlation function for day 6 was negative. Based on this evidence, the average temperature data was accepted as uncorrelated.

Frequency histograms were developed from the data. An example of these is day 11 which is shown in Figure 10. Only the normal distribution appeared to fit the shapes of the histograms and accepts the range of possible data.

Estimation for normal distribution parameters was very simple using maximum likelihood (Appendix A-1). These parameter estimates are displayed in Table 8.

The Kolmogorov-Smirnov goodness-of-fit test and the critical values for normal distribution were used to establish the fit. These test results are displayed in Table 8. A normal distribution was found to fit all 16 data sets at the five percent significance level.

The average temperature model was validated for the normal distribution. Cumulative probabilities of 1975 data varied from .082 to .967. Thus all of the data were accepted by the model and produced reasonable cumulative probabilities. The model is considered to be valid.

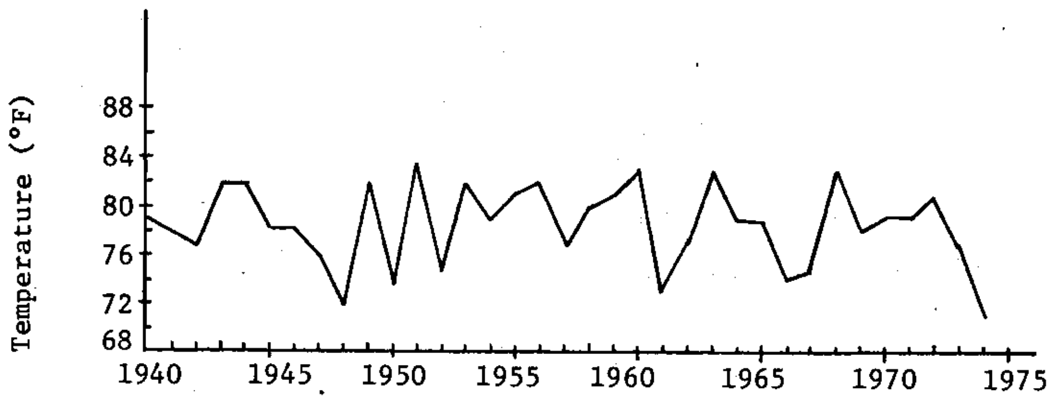

Figure 9. Average Temperature (Day 10)

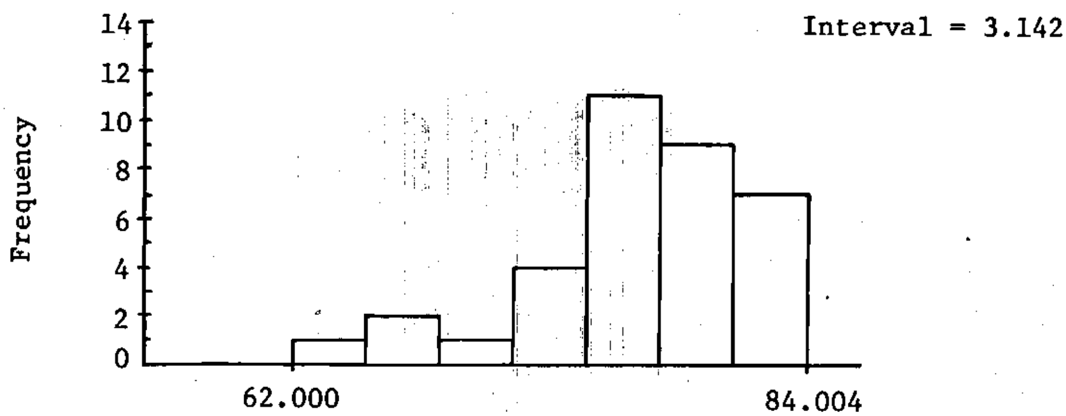

Figure 10. Average Temperature (Day 11)

Table 8. Average Temperature Models

Normal Distribution			
Day	Mean	Variance	Test Statistic**
1.	41.4	140.19	.118
2.	42.3	80.65	.101
3.	48.9	73.85	.102
4.	51.8	68.56	.090
5.	59.2	49.93	.106
6.	67.4	44.29	.093
7.	73.7	32.83	.072
8.	77.5	13.45	.076
9.	77.4	18.13	.128
10.	78.4	11.38	.072
11.	76.7	23.29	.077
12.	71.3	29.11	.083
13.	64.9	35.59	.075
14.	53.6	38.52	.090
15.	44.8	57.11	.085
16.	44.8	84.39	.105

* Fails to fit

** Critical value = (α=.05) = .150

Barometric Pressure

Thirty-three years of data were graphed for each data set. All the data varied over a very narrow range. As a result of this range, the variability found in days 1, 2, 10, 11, and 12 was not extreme (Figure 11). The remaining 11 days appeared to be stable.

The two data sets chosen for additional analysis were days 3 and 14. Auto-correlation and partial auto-correlation functions were calculated for both days. The results for day 3 were inconclusive. A clear correlation was found in data for day 14 (Figure 11). The characteristics of the correlation functions indicated a one parameter moving average model at a lag of two. Because of the strength of these results, all barometric pressure data sets were checked for correlations. The purpose of this additional analysis was to determine if this was a characteristic of all of the data sets or an isolated instance. Days 4, 6, 7, 9, 13, 15 and 16 did not have any discernable structure. The remaining days could be considered to have had structure only when the level of significance was relaxed. This was done by calculating the significant value of the partial auto-correlation functions for a sample size of 60. It is difficult to explain the results for day 14. They could be caused by true correlations which could also be present in the other data sets but were not found because of small sample size. Day 14 could be an exception, which would not be present if more data had been used. Based on the overall results of the 16 days, it was concluded that the data were not correlated and would therefore be suitable for modeling.

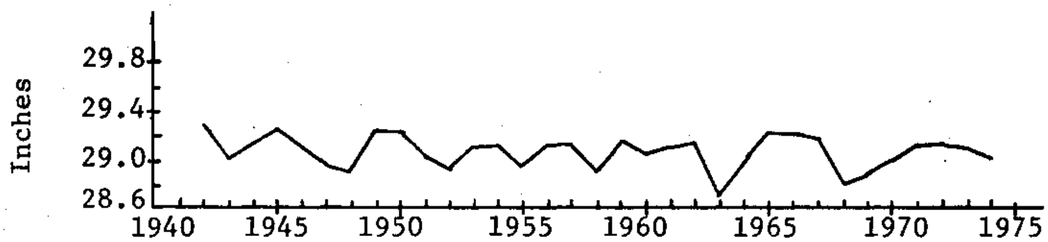

Figure 11. Barometric Pressure (Day 14)

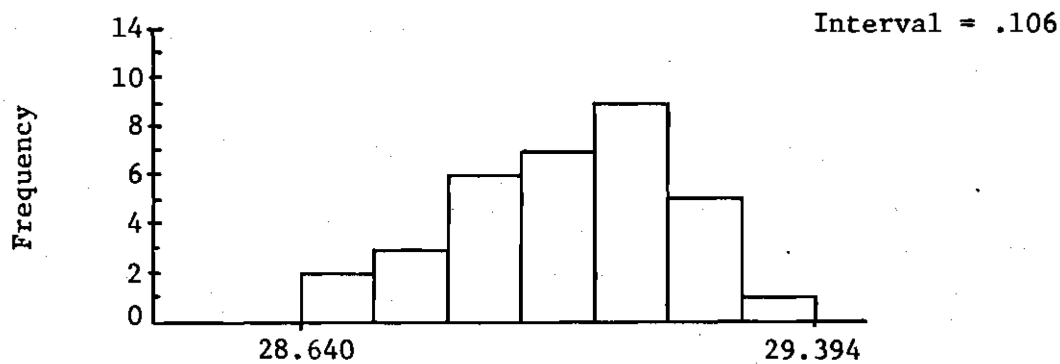

Figure 12. Barometric Pressure (Day 3)

Frequency histograms were developed for the 16 data sets. A normal distribution would appear to fit the shapes of the histograms (Figure 12). The data ranged between 28.00 and 30.00 inches which would cause the distributions to be compact and thus make the probabilities of zero or negative values effectively zero. Other distributions such as the gamma were not acceptable because of the presence of unacceptable tails that would not accurately represent such compact data. Therefore the normal distribtuion was the only distribution tested for modeling barometric pressure.

Maximum likelihood estimators were used to calculate the means and variances of the normal distributions (Appendix A-1). These parameters are presented in Table 9. The low values of the variances were an indication of the compactness of the data. These parameter estimates were used in testing for distribution fit.

The Kolmogorov-Smirnov goodness-of-fit tests were conducted using the critical values calculated for the normal distribution. Of the 16 models only one, day 9, failed to fit the data. One failure out of 16 was not significant.

The normal distribution was validated for the 15 days that showed a fit. The sixteenth, day 9, was not included in the validation effort because the normal distribution did not fit the data. Of these 15 days, only day 12 was not validated using the 1975 data. Day 12 had a cumulative probability of .000 for an observed value of 28.45 inches. The smallest data value used in the estimation of the parameters of day 12 was 28.62. This problem could be corrected with the addition of more data. The models can be considered valid.

Table 9. Barometric Pressure Models

Normal Distribution			
Day	Mean	Variance	Test Statistic**
1	29.09	.040	.091
2	29.10	.042	.070
3	29.03	.028	.047
4	28.98	.029	.090
5	29.03	.033	.094
6	29.01	.016	.071
7	28.99	.008	.080
8	28.97	.012	.127
9	29.04	.032	.239*
10	29.01	.008	.069
11	29.03	.006	.068
12	29.03	.017	.088
13	29.03	.014	.130
14	29.08	.018	.134
15	29.15	.041	.120
16	29.02	.051	.123

* Fails to Fit

** Critical value for normal distribution ($N=38$, $\alpha=.05$) is 0.154.

Relative Humidity

Thirty-three data points for each data set were graphed against time. Five of these days, 7, 8, 9, 10, and 11 showed small degrees of fluctuations (Figure 13). These days span the period June 7 to September 3 inclusive. The data for the other 10 days showed considerable variability between the values 10 to 100 percent.

Three days were chosen for the correlation tests. These were days 2, 9, and 10. Three days, rather than two, were tested because the data had stronger indications of serial correlations. A graph of day 10 is shown in Figure 13. Tests on the three data sets did not not show correlations

Frequency histograms tend to show at least a slight negative skew. Days 1, 2, 4, and 9 were the exceptions. These histograms were either flat or unimodal. All of the data were between zero and one which suggested a beta distribution. Because the beta distribution was particularly suited to bounded data and was capable of assuming many shapes, it was selected as the candidate distribution.

Two methods were used to estimate the parameters of the beta distribtuion. Estimation formulas described by Yao (45) were first used because of their simplicity. These estimators were subsequently re-estimated using maximum likelihood (Appendix A-4). In some cases these new estimators proved to be worse than the first sets. Finally the best parameter set of the two methods was used in the models.

Critical values were not available for Kolmogorov-Smirnov goodness-of-fit tests of beta distributions. Determination of the fit

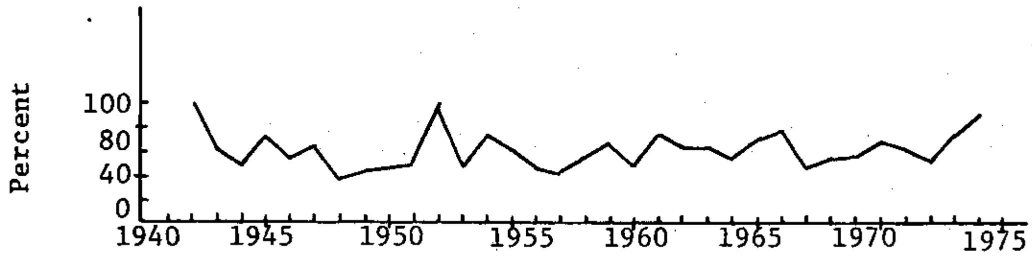

Figure 13. Relative Humidity (Day 10)

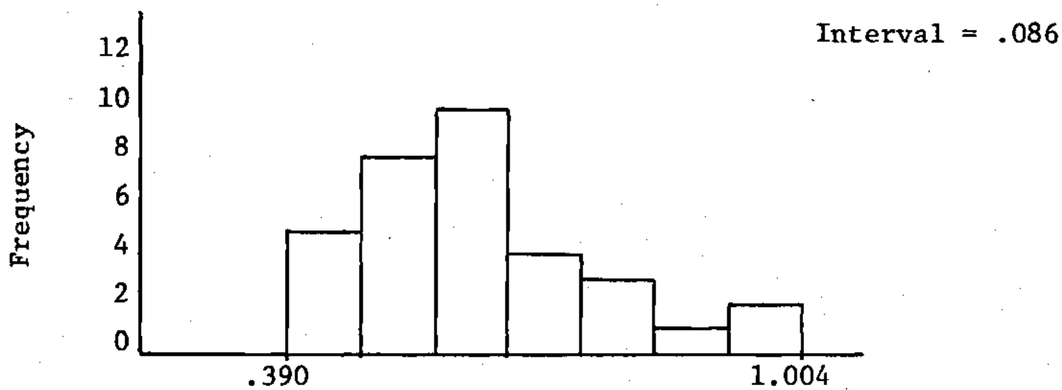

Figure 14. Relative Humidity (Day 10)

was a combination of the obvious supplemented by a chi-square goodness-of-fit test. None of the results was small enough for immediate acceptance. Test statistics that were obviously too large to establish fit were accepted. Doubtful tests were conducted again using the chi-square goodness-of-fit test. These results are presented in Table 10. Only seven of the 16 days were fitted by the beta distribution. This was not a significant result.

Therefore, based on these tests, the beta distribution is not a suitable model of daily relative humidity.

Table 10. Relative Humidity Models

Beta Distribution			
Day	p	q	Test Statistic**
1.	1.93	1.20	.347*
2	1.73	1.64	.158
3.	2.72	2.36	.135
4.	1.88	1.40	.201
5.	1.92	1.76	.120
6.	2.08	2.08	.303*
7.	5.21	4.92	.113
8.	5.32	4.15	.333*
9.	10.40	8.30	.387*
10.	6.02	3.82	.549*
11.	5.25	4.54	.182
12.	4.65	3.88	.249*
13.	3.32	3.14	.143
14.	2.52	2.76	.233*
15.	2.14	2.18	.238*
16.	2.42	1.56	.420*

* Failed to pass

** Fit was determined using a chi-square goodness-of-fit test at a 5% significance level.

Cloud Cover

The graphs of tenths of sky cover, sunrise to sunset, showed very erratic behavior (Figure 15). This was true of all 16 days. Using these graphs, two days were selected for analysis for correlations.

Days 1 and 2 were analyzed for trends and serial correlations. Day 1 showed a very weak correlation. Day 2 did not exhibit any correlation. Based on these results, cloud cover data could not be excluded from this study because of correlation.

Frequency histograms showed three different shapes. Days 1, 2, 3, 6, and 16 were bimodal with the high points at each end (Figure 16). Days 9 and 14 were approximately flat. The remaining days were unimodal. This was considerable variability for the same type of data. Considering the range of the data, zero to one, and the variety of shapes, the beta distribution was the natural choice for modeling.

Problems in parameter estimation paralleled those encountered with the relative humidity data. The same estimators were used and a choice was made between the two sets. The parameter set that best fit the data in the goodness-of-fit test was the one chosen.

The goodness-of-fit was determined in the same manner as relative humidity. Only day 11 had a test result less than 1.00 (Table 11) and was accepted without further tests. Six additional days were found to be fitted by the beta distribution. A chi-square test was used to establish a fit for days 2, 6, 7, 12, 14, and 16. Thus only seven of 16 days were fitted by the beta distribution. This was not enough successes to accept the beta distribution as an appropriate model of cloud cover.

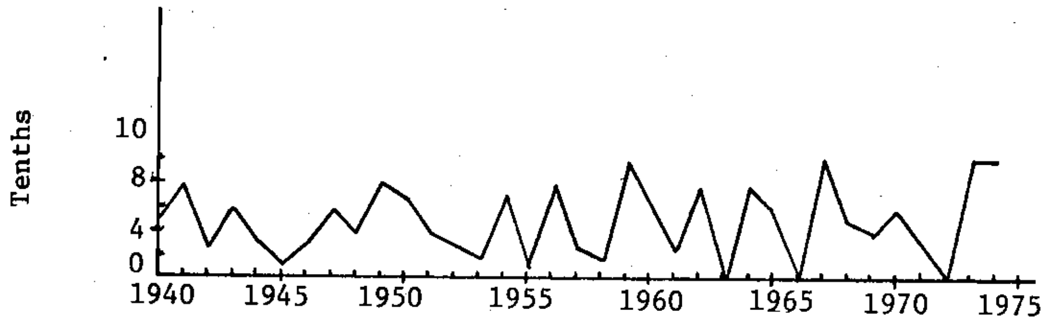

Figure 15. Cloud Cover (Day 7)

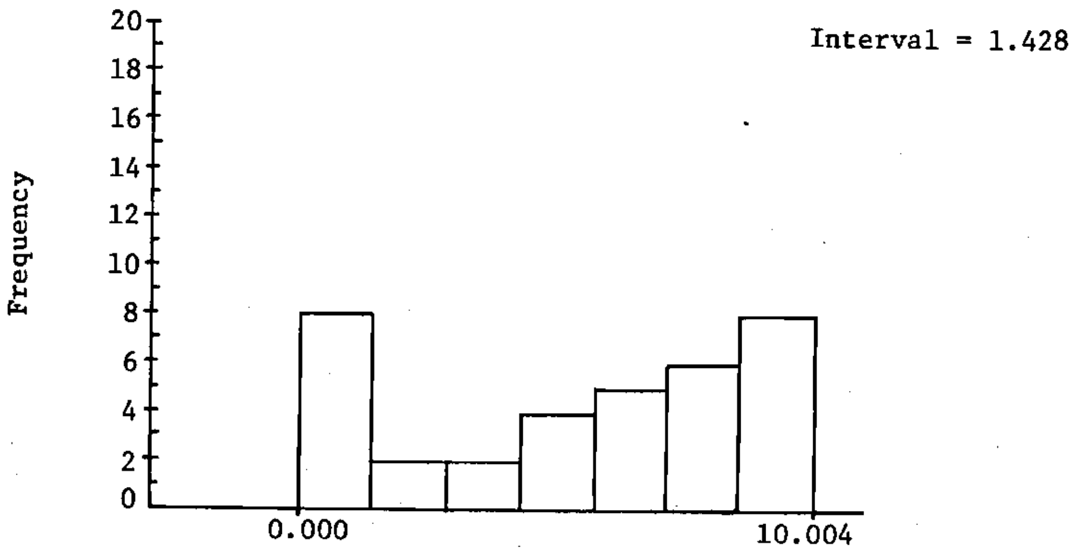

Figure 16. Cloud Cover (Day 6)

Table 11. Cloud Cover Models

Beta Distribution			
Day	p	q	Test Statistic**
1.	.42	.30	.174*
2.	.25	.24	.143
3.	.27	.18	.217*
4.	.53	.23	.402*
5.	.43	.30	.202*
6.	.54	.43	.160
7.	.88	.90	.108
8.	1.05	.70	.232*
9.	1.14	.76	.228*
10.	1.35	.92	.272*
11.	.77	.76	.096
12.	.50	.60	.190
13.	.26	.63	.513*
14.	.51	.52	.171
15.	.22	.30	.343*
16.	.57	.41	.159

* Fails to fit

** Fit was determined using a chi-square goodness-of-fit test at a 5% significance level.

Thunderstorm or Distant Lightning

A thunderstorm or lightning sighting was reported as an event if it was seen or heard. The daily data only shows that either none or at least one event occurred on that day. It would do little good to graph or develop frequency histograms of these data. This is as a zero or one situation that was best described by a simple probability of occurrence.

These probabilities were developed from the 35 data points for each day. Division of the number of successes by 35 produced the probabilities shown in Table 12. This result is the probability of at least one occurrence of a thunderstorm or distant lightning within line of sight over 360 degrees.

Validation of the probabilities was accomplished using 1975 thunderstorm events. Thunderstorms or distant lightning occured on three of the 16 days in 1975. Days 1, 8, and 9 had thunderstorms which the model predicted with the probabilities of .0571, .4857, and .2286 respectively. There were no problems with these probabilities and thus the model can be considered valid.

Table 12. Thunderstorm or Distant Lightning Models

Day	Probabilities
1.	.0571
2.	.0571
3.	.0571
4.	.2000
5.	.0857
6.	.2000
7.	.2571
8.	.4857
9.	.2286
10.	.4286
11.	.1714
12.	.0281
13.	.0571
14.	.0000
15.	.0286
16.	.0286

Sunshine

Graphs of all 16 data sets show considerable fluctuation. The data for this measure vary between zero and one (Figure 17). This was seen in all graphs except day 7 (June 1). Twenty percent was the 35 year lower bound for day 7. Only six of these days showed any signs of trend or correlation.

Days 2 and 11 were chosen for the correlation function analysis. Day 2 (Figure 17) showed a possible correlation at lag three. In contrast day 11 had no discernible structure. Given the two contrasting results, it was concluded that there was insufficient evidence of correlation to exclude percent sunlight from the study.

The frequency histograms were unimodal and bimodal. Days 1, 2, 3, 4, 5, 15, and 16 were bimodal with the high points at both ends (Figure 18). The remaining nine days were unimodal. These shapes and the range of the data, zero to one, would require a beta distribution. A beta distribution can be used to model data that is bounded on both ends. In addition the beta can assume several different shapes depending on the relative magnitudes of the distribution parameters.

Estimation of the parameters was complicated by the same problems encountered with relative humidity and cloud cover. Both estimation methods shown in Appendix A-4 were required. The parameters that gave the best results in the Kolmogorov-Smirnov goodness-of-fit test were the ones presented in Table 13.

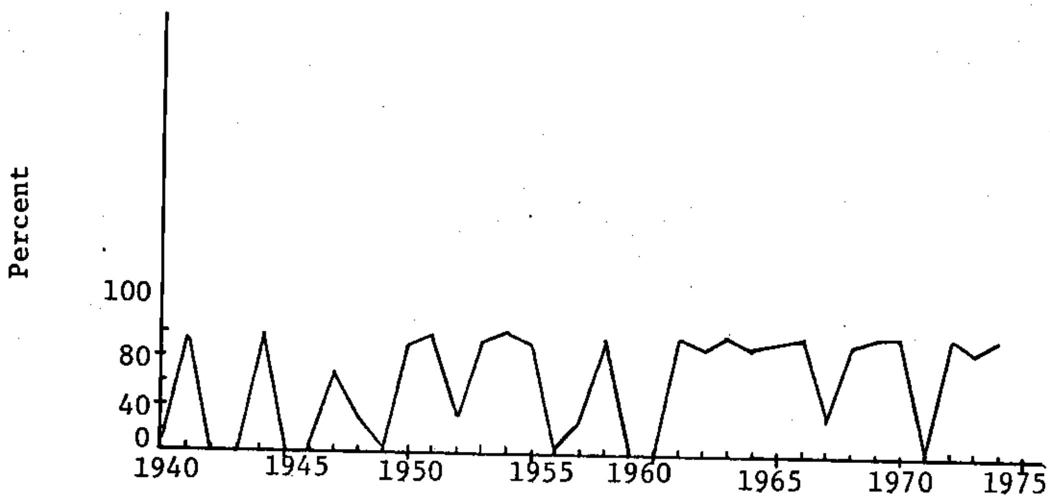

Figure 17. Sunshine (Day 2)

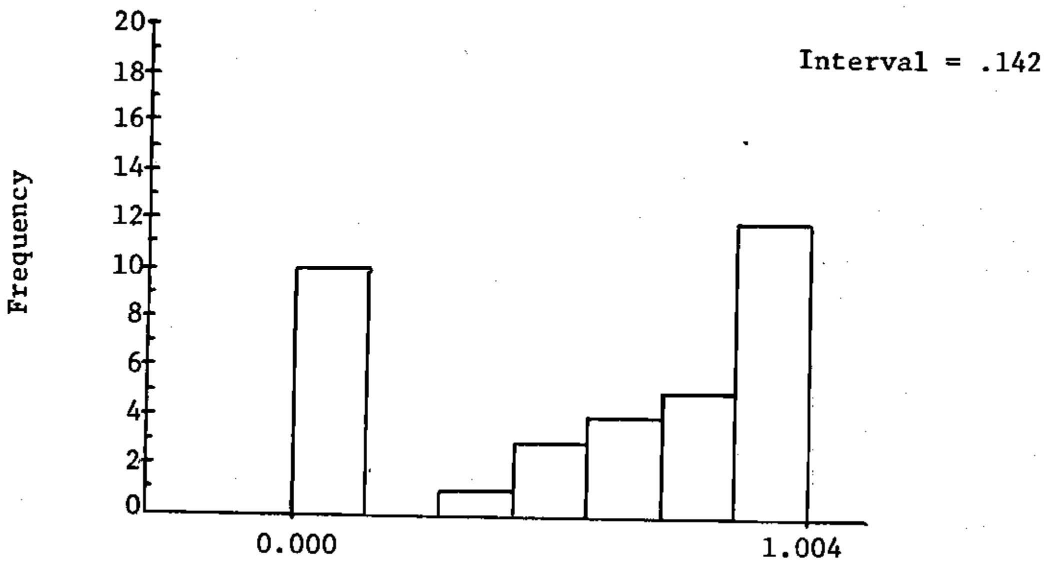

Figure 18. Sunshine (Day 1)

Table 13. Sunshine Models

Beta Distribution			
Day			Test Statistic**
1.	.14	.13	.200
2.	.11	.10	.257*
3.	.50	.38	.199
4.	.26	.30	.200
5.	.43	.27	.287*
6.	.22	.13	.375*
7.	1.49	.48	.613*
8.	.47	.26	.501*
9.	.51	.35	.355*
10.	1.06	.70	.277*
11.	.29	.17	.406*
12.	.31	.19	.423*
13.	.25	.09	.598*
14.	.64	.31	.425*
15.	.16	.09	.341*
16.	.42	.40	.171*

* Fails to fit

** Fit was determined using a chi-square goodness-of-fit test at a 5% significance level

None of the test statistics developed by the Kolmogorov-Smirnov goodness-of-fit test were small enough to allow acceptance. Chi-square goodness-of-fit tests were used to establish distribution fit to the data for days 1, 3, 4, and 16. The remainder of the data sets was not satisfactorly fitted by the beta distribution. Four successes out of 16 trys was not sufficient proof that the beta distribution was the correct model for sunshine.

CHAPTER IV

RESULTS AND CONCLUSIONS

The results and conclusions of the analysis conducted in chapter III are presented for each one of the climatological measures. Nine of the climatological models were successful. Three failed that might have succeeded if more data had been available. The major shortcoming of this study is the relatively small amount of good data that was available for developing and testing the models.

Precipitation

It was determined that daily precipitation data have a relatively large number of zero values. The most applicable distributions cannot use zero values. Enough additional data must be collected to total 35 or more positive values that could be used to estimate stable parameters. The effect of a larger time span is to eliminate zero values by combining data from more than one day. Given these additional data it is highly probably that models of daily precipitation can be developed that would provide cumulative probabilities of precipitation amounts.

In the absence of additional data, the probabilities of precipitation provided in this study should provide information of value in planning. This information would be the probability of more than a trace amount of precipitation. These probabilities can be refined by the addition of each new annual unit data as it becomes available.

Average Wind Speed

Two distributions fit the average wind speed data. Both the gamma and lognormal distributions were used to model average wind speed. Both distributions were successful and can be used to model daily average wind speed.

Maximum Wind Speed

Maximum wind speed can be modeled by either the gamma or type II extreme value distributions. Both of these distributions successfully fit the data in 15 of 16 cases. The type II extreme value distribution may have a slight advantage because of the ease in calculating the cumulative probabilities. Its cumulative distribution function does not involve the evaluation of integrals which is the case with the gamma. Such a choice is left to the user and the particular use.

Wind Direction

The probabilities of wind direction are a viable means of predicting wind directions. It would have been more accurate to divide the compass into 16 equal sections, but this was not possible with the way the data are reported. One problem that needs to be remedied is the presence of zero probabilities for some directions. Given only 35 data points to be divided among 16 directions, it is highly probable that zero probabilities will occur. Addition of more data would probably result in these zeros changing to a small probability. The result would be a more realistic model that would be more accurate.

Maximum Temperature

The normal and three quarters of the type I extreme value distributions were found suitable models of maximum temperature. Unfortunately three quarters of a distribution is the same as no distribution when applied to a prototype. The inability of the type I extreme value distribution to model the summer from day 8 (June 21) to day 11 (September 3) must be resolved. This can be done first by adding more data. Second, the days near the four failures could be modeled. The results of these additional models could help determine if these are isolated cases. If not, then the distribution is not suitable for modeling the entire year. Until such time as this could be accomplished, the normal is the proper choice for modeling daily maximum temperature.

Minimum Temperature

The results obtained for minimum temperature are very similar to those of maximum temperature. The normal distribution has been found to be a valid model for daily minimum temperature. A type I extreme value was capable of fitting only 12 of the 16 days. Three of the failures were in December and January which reduces the validity of the model. The required solution is the same as for maximum temperature. More data are required and additional modeling of selected days are necessary before the type I extreme value distribution can be accepted or rejected for modeling minimum temperature.

Average Temperature

Only the normal distribution was considered for modeling average temperature. The results of the goodness-of-fit tests were 100 percent positive. Thus the normal distribution can be expected to provide valid results. It is reasonable to expect these models to improve as more data are used in the estimation of parameters.

Barometric Pressure

The normal distribution was found to be a fit for 15 of 16 data sets. These data used in constructing the models showed relatively small variability. This resulted in compact distributions that can provide probabilities for a small range of values. Day 12 is an example one such distribtution that was too compact. A validation data value of 28.45 inches was less than the minimum value of 28.62 inches used in constructing the normal distribution model for day 12. It is reasonable to expect similar failures with the models as they are at present. Therefore the addition of more data for the improvement of these models is necessary before they can be considered operational.

Relative Humidity

The beta distribution fit the relative humidity data for 7 of 16 days. This was not significant. Therefore the beta distribution was not accepted as a model for relative humidity. Yao (45) used 53 years of data in the Washington, D.C. study as compared to the 33 years of data available for this model. Yao's results were arrived at graphically.

It would be reasonable to expect data to smooth as the sample size is increased which could result in improved parameter estimation and fit. Therefore more data would have to be added to the model before the beta distribution could be fully accepted or rejected.

Cloud Cover

The same problems were encountered in developing a cloud cover model. Only seven of the 16 days were fitted by the beta distribution. Cloud cover data vary between zero and one as does relative humidity. A difference between the two is the continuous nature of relative humidity data versus the discreet characteristics of cloud cover data. Cloud cover is measured in tenths and thus can only assume 11 values, 0.0, 0.1, 0.2, ... 1.0. It is uncertain whether this would have an effect on the results. Additional data and more analysis would be required before the beta distribution could be judged suitable or unsuitable as a model.

Thunderstorm or Distant Lightning

The method developed for modeling thunderstorm data is probably best suited for the data. Another, more advanced technique, could be used to model the number of such events each day, but that is not the nature of the available data. This model could also be improved with the addition of more data. It would be reasonable to find at least some of the zero probabilities assuming a small positive value as the samples size was increased.

Sunshine

The modeling of sunshine data was not successful. A beta distribution was capable of fitting only four of the 16 days. This was a problem shared with the relative humidity and cloud cover data sets. Additional data is required before the beta distribution can be ruled out as a good candidate for modeling sunshine. A valid beta distribution model of the percent of daily sunshine, coupled with a table showing the sunshine hours possible, would provide a good predictor of daily sunshine.

Systems Approach

This study was directed toward the prediction of daily climate using a system of 12 climatological measures. Nine of the 12 measures were successfully modeled in this study. The results for the other three were inconclusive rather than negative. In a practical sense this actually represents more than 75 percent of the necessary information. The major factors of climate were modeled. These factors were precipitation, temperature, wind, and thunderstorms. Using these models, it is possible to make improved predictions for sunlight, cloud cover, and relative humidity. Thus, the models developed in this study can be used to predict a significant portion of daily climate.

CHAPTER V

RECOMMENDATIONS

Additional data should be collected for all of the climatological measures before the results are extended to the remaining days of the year. This is especially necessary for the establishment of valid models of precipitation, relative humidity, cloud cover and sunshine. More data would probably result in the extreme value distributions being validated for maximum and minimum temperature. Barometric pressure models would be improved and would be less likely to fail for larger or smaller data values.

Once these data have been accumulated and used to improve the models, validation can be reattempted. If validation of these improved models is successful, the next step is to extend the model to the remaining days of the year for Atlanta, Georgia.

If the prototype model is successfully extended to the remaining days of Atlanta's climate, the next step is to investigate the methodology for extreme climates. Some of the extremes of climate would be desert, artic, tropical rain forest, mountain, and others. The results of modeling extreme climates could be compared to the moderate climate of Atlanta, Georgia.

Changes should be made in the reporting of wind direction data. Reporting by tens of degrees or one of the 16 compass points is inaccurate. More accurate wind direction data could be used to develop models without the bias caused by the reporting method.

Bivariate methods could be used to extend the models of wind speed and direction. Such models would join together two measures that are strongly related. The resulting model would be an improvement in predicting wind and its attributes.

Improved tables of Kolmogorov-Smirnov critical values should be prepared for all of the distributions used in modeling. The Kolmogorov-Smirnov test would then more useful and accurate. It would also reduce the missuse of the non-parametric tables if the parametric tables were widely available and known.

APPENDICES

APPENDIX A

1. Normal Distribution

There are two parameters associated with the normal distribution, the mean (μ) and standard deviation (σ). Both of these parameters are estimated from the sample data and are approximated by $\bar{x}$ and s respectively. (19)

Probability Density Function

$$f[x] = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})} \qquad -\infty < x < +\infty$$

Cumulative Distribution Function (Standarized)

$$Pr[\gamma \leq y] = \frac{1}{\sqrt{2\ \pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}x^2} dx$$

$$\text{where} \quad y = (\frac{x - \mu}{\sigma})$$

Parameter Estimation (Maximum Likelihood)

$$\bar{x} = n^{-1} \sum_{j=1}^{n} x_j$$

$$S^2 = n^{-1} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

2. Lognormal Distribution

The lognormal distribution is a transformation of the normal distribution with mean and standard deviation. These parameters are estimated from the natural logarithmic transform of the data by the parameters ζ and σ respectively. (19)

$$Z_j = \log_e (x_j)$$

Probability Density Function

$$f(x) = \frac{1}{x\,\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\log(x)-\zeta)^2}{\sigma}} \quad x > 0$$

Cumulative Distribution Function

None given

Parameter Estimation (Maximum Likelihood)

$$\zeta = \bar{Z} = n^{-1} \sum_{j=1}^{n} Z_j$$

$$\sigma = [n^{-1} \sum_{j=1}^{n} (Z_j - \bar{Z})^2]^{\frac{1}{2}}$$

3. Gamma Distribution

The gamma is a two-parameter distribtuion. A scale parameter, β, and a shape parameter, γ, are estimated from the data. (38)

Probability Density Function

$$f(x) = \frac{1}{\beta^{\gamma}\,\Gamma(\gamma)} x^{\gamma-1} e^{-x/\beta} \quad \begin{matrix} \beta > 0 \\ \gamma > 0 \end{matrix}$$

Cumulative Distribution Function

None given

Parameter Estimates (Maximum Likelihood)

$$A = \log_e \bar{x} - \frac{1}{n} \Sigma \log_e x$$

$$\gamma = \frac{1 + \sqrt{1 + 4\ A/3}}{4A}$$

$$\bar{x}/\beta \times \gamma \doteq o$$

4. Beta Distribtuion

The beta distribution requires four parameters. Two parameters pertain to the bounds of the data values and are this study, a=o and b=1. This reduces the beta distribtuion to to two parameters p and q. (19)

Probability Density Function

$$f(x) = \frac{1}{B(p.q)} x^{p-1} (1-x)^{q-1} \quad (o \leq x \leq 1)$$

Cumulative Distribution Function

$$Pr[X \leq x] = I_x \ (p,q) = \frac{1}{B(p,q)} \int_0^x t^{P-1} (1-t)^{q-1} dt$$

where $B(p,q) = \Gamma(P)\ (q) / \ \Gamma(P+q)$

Parameter Estimates (Method of Moments) (41)

$$p = \frac{\mu_1'(\mu_1' - \mu_2')}{\mu_2' - \mu_2'^2} \qquad p = \frac{(1-\mu_1')(\mu_1' - \mu_2')}{\mu_2' - \mu_1'^2}$$

$$\text{where} \quad \mu_1' = n^{-1} \sum_{j=1}^{n} x_j$$

$$\mu_2' = [n^{-1} \sum_{j=1}^{n} x_j^2]^{\frac{1}{2}}$$

Parameter Estimates (Maximum Likelihood) (19)

$$\psi(\hat{p}) - \psi(\hat{p} + \hat{g}) = n^{-1} \sum_{j=1}^{n} \log_e \left(\frac{\psi j - a}{b - a}\right)$$

$$\psi(\hat{g}) - \psi(\hat{p} + \hat{g}) = n^{-1} \sum_{j=1}^{n} \log_e \left(\frac{b - \psi j}{b - a}\right)$$

where $b=1$, $a=0$

ψ (.) is the digamma function. These are solve iteratively for solutions to $\hat{p}$ and $\hat{g}$.

5. Fisher Tippet Type I Extreme Value Distribution

The type I extreme value distribution involves two parameters. It can be used to model both maximum or minimum data. Only one set of maximum likelihood estimators are given in Johnson and Kotz (19) for both maximum and minimum data. Gumbel (16) provides different method of moment estimators for maximum and minimum values. In addition the cumulative distribution functions are different for maximum and minimum value distribution. Theta is the scale parameter and psi (ζ) is the location parameter.

Probability Density Function

$$f(x) = \frac{e^{-(x-\zeta)-e^{-(x-\zeta)}}}{b} \qquad -\infty < x < +\infty$$

Cumulative Distribution Function

$$\Pr[X \leq x] = \exp\{-e^{-(x-\zeta)/\theta}\} \qquad -\infty < x < +\infty$$

Parameter Estimators (Maximum Likelihood) (19)

$$\hat{\theta} = n^{-1} \sum_{j=1}^{n} x_j - [\sum_{j=1}^{n} x_j e^{-x_j/\hat{\theta}}] [\sum_{j=1}^{n} e^{-x_j/\hat{\theta}}]^{-1}$$

$$\hat{\zeta} = -\hat{\theta} \log [n^{-1} \sum_{j=1}^{n} e^{-x_j/\hat{\theta}}]$$

Parameter estimators (Method of Moments) (16)

$$\hat{\theta} = \frac{\sqrt{6}}{\pi} s(x) \qquad s(x) = \left[\sum_{j=1}^{n} \frac{(x_j - \bar{x})^2}{n}\right]^{\frac{1}{2}}$$

$$\hat{\zeta} = \bar{x} \quad \gamma\hat{\theta} \qquad \gamma = 35 = .5403$$

Minimum Value(15)

Probability Density Function

None Given

Cumulative Distribution Function

$$\Pr[X \leq x] = 1 - \exp\{-e^{(x-\zeta)/\theta}\}$$

Parameter Estimators (Method of Moments)

$$\hat{\theta} = \frac{\sqrt{6}}{\pi} \; s(x)$$

$$\hat{\zeta} = \bar{x} + \gamma\hat{\theta}$$

6. Fisher-Tippett Type II Extreme Value Distribution

The Fisher-Tippett Type II distribution involved three parameters. Psi (ζ) is a lower bound that is equal to zero for this study. Theta (θ) is a scale parameter and Kappa (κ) is a shape parameter. The data must all be greater than zero. (37)

Cumulative Distribution Function

$$\Pr(X \leq x) = \begin{cases} 0 & (x < \zeta) \\ \exp -\{(\frac{x-\zeta}{\theta})^{-\kappa}\} & (x \geq \zeta) \end{cases}$$

Parameter Estimators (Maximum Likelihood)

$$\hat{\kappa} = \left[-n \quad - \sum_{j=1}^{n} \log x_j + n \sum_{j=1}^{n} \frac{x_j^{-\kappa} \log x_j}{\sum_{j=1}^{n} x_j^{-\hat{\kappa}}} \right]^{-1}$$

$$\hat{\theta} = \left[\frac{n}{\sum_{j=1}^{n} x_j^{-\hat{\kappa}}} \right]^{1/\kappa} \qquad \theta > 0$$

BIBLIOGRAPHY

1. Angell, J. K., and J. Korshover, "Variation in Sunshine Duration Over the Contiguous United States Between 1950 and 1972," Journal of Applied Meteorology, Vol. 11 (September 1975), pp. 1174-1181.

2. Barger, Gerald L., and H. C. S. Thom. "Evaluation of Drought Hazard," Agronomy Journal, Vol. 41, No. 11 (November 1949), pp. 519-526.

3. Box, George E. P., and Gwilyn M. Jenkins. Time Series Analysis Forecasting and Control, Holden-Day Inc., San Francisco, Ca., 1970.

4. Bridges, T. C., and C. T. Haan. "Reliability of Precipitation Probabilities Estimated From the Gamma Distribution," Monthly Weather Review, Vol. 100, No. 8 (August 1972), pp. 607-611.

5. Computer Programs for the Analysis of Univariate Time Series Using the Methods of Box Jenkins, Industrial Engineering Computer Library, Georgia Institute of Technology, Atlanta, Georgia, 1977.

6. Court, Arnold, "Wind Roses," Weather, Vol. 18. (April 1963), pp. 106-110.

7. Crutcher, Harold L. "A Note on the Possible Misuse of the Kolmogorov-Smirnov Test," Journal of Applied Meteorology, Vol. 14 (December 1975), pp. 1066-1603.

8. Deutch, S. J. Unpublished Notes, Georgia Institute of Technology, Atlanta, Georgia, 1974.

9. DiCarlo, Vincent. Unpublished Interview, United States Weather Bureau, Atlanta, Georgia, 1976.

10. Duncan, Acheson, J. Quality Control and Industrial Statistics, 2nd Edition, Richard D. Irwin, Inc., Homewood, Ill., 1959.

11. Encyclopedia Britannica, The New, 15th Edition, Chicago, Ill., Vol. 3 and 9.

12. Falls, Lee E., William O. Williford, and Michael C. Carter. "Probability Distributions for Thunderstorm Activity at Cape Kennedy, Florida," Journal of Applied Meteorology, Vol. 10 (February 1971).

13. Falls, Lee W., "The Beta Distribution: A Statistical Model for World Cloud Cover," NASA Technical Memorandum TMX-64714, Marshall Space Flight Center, Alabama, 1973.

14. Friedman, D. G., and B. E. Janes. "Estimation of Rainfall Probabilities," Bulletin No. 332, Connecticut Agricultural Experiment Station, Storrs, Connecticut.

15. Garrison, David D. A Pilot Study of the Probability of Occurrence of Climatological Phenomena in the Atlanta Area, Unpublished Report, Georgia Institute of Technology, Atlanta, Georgia, 1965.

16. Gumbel, Emil J., Statistics of Extremes, Columbia Univeristy Press, New York, New York, 1958.

17. Hines, William W., and Douglas C. Montgomery. Probability and Statistics in Engineering and Management Science, The Ronald Press Company, New York, 1972.

18. Johnson, C. G. Unpublished Notes, Georgia Institute of Technology, Atlanta, Georgia, 1975.

19. Johnson, Normal L., and Samuel Kotz. Continuous Univariate Distributions - 1, Houghton Mifflin Company, Boston, 1970.

20. Johnson, Norman L., and Samuel Kotz. Continous Univariate Distributions - 2, Houghton Mifflin Company, Boston, 1970.

21. Local Climatological Data, National Oceanic and Atmospheric Administration, U. S. Department of Commerce, 1950 - 1977.

22. Luna, R. E., and H. W. Church. "Estimation of Long-Term Concentrations Using a 'Universal' Wind Speed Distribution," Journal of Applied Meteorology, Vol. 13 (December 1974), pp. 910-916.

23. Madden, R. W., and W. Sudeh. "Empirical Estimates of the Standard Error of Time-Averaged Climatic Means," Journal of Applied Meteorology, Vol. 14 (March 1975).

24. Massey, Frank J. Jr. "The Komogorov-Smirnov Test for Goodness of Fit," American Statistical Association Journal (March 1951), pp. 68-78.

25. Meteorological Records of Surface Observations, National Weather Service, Atlanta, Georgia.

26. Mielke, Paul W., Jr. "Another Family of Distributions for Describing and Analyzing Precipitation Data," Journal of Applied Meteorology, Vol. 12 (March 1973), pp. 275-280.

27. Mooley, D. A. "Independence of Monthly and Bimonthly Rainfall Over Southeast Asia During the Summer Monsoon Season," Monthly Weather Review, Vol. 98, No. 2 (February 1970), pp. 154-160.

28. Mooley, D. A., and G. Appa Rao. "Distribution Function for Seasonal and Annual Rainfall Over India," Monthly Weather Review, Vol. 99, No. 10 (October 1971), pp. 796-799.

29. Mooley, D. A., "Gamma Distribution Probability Model for Asian Summer Monsoon Monthly Rainfall," Monthly Weather Review, Vol. 101, No. 2 (February 1973), pp. 160-175.

30. Mooley, D. A., "An Estimate of the Distribution and Stability Period of the Parameters of the Gamma Probability Model Applied to Monthly Rainfall Over Southeast Asia During the Summer Monsoon," Monthly Weather Review, Vol. 101, No. 12 (December 1973), pp. 884-890.

31. Okulaja, F. Ola. "The Frequency Distribution of Lagos/Ikeja Wind Gusts," Journal of Applied Meteorology, Vol. 7 (June 1968). pp. 379-383.

32. Ostle, Bernard, Statistics in Research, 2nd Edition, Iowa State University Press, Ames, Iowa, 1966.

33. Sakamoto, Clarence M. "Application of the Poisson and Negative Binomial Models to Thunderstorm and Hail Days Probabilities in Nevada," Monthly Weather Review, Vol. 101, No. 4 (April 1973), pp. 350-355.

34. Shenton, L. R., and K. O. Bowman. "Remarks on Thom's Estimators for the Gamma Distribution," Monthly Weather Review, Vol. 98, No. 2 (February 1970), pp. 154-160.

35. Shenton, L. R., and K. O. Bowman. "Note on the Sample Size to Achieve Normality for Estimators for the Gamma Distribution," Monthly Weather Review, Vol. 101, No. 12 (December 1973), pp. 891-892.

36. Skees, P., and L. R. Shenton. "Comments on the Statistical Distributions of Rainfall Per Period Under Various Transformations," Proceedings of the Symposium on Statistical Hydrology, USDA Misc. Publ. 1275, 1974, pp. 172-196.

37. Thom, H. C. S., "Frequency of Maximum Wind Speeds," Proceedings of American Society of Civil Engineers, Vol. 80 (November 1954).

38. Thom, H. C. S., "A Note on the Gamma Distribution," Monthly Weather Review, Vol. 86, No. 4 (April 1958), pp. 117-122.

39. Thom, H. C. S., "Standard Deviation of Monthly Average Temperature," EDS-3, Environmental Data Service, Silver Spring, Maryland, 1968.

40. Thom, H. C. S., "Climatic Probability for Favorable Viewing Conditions Along Path of March 7, 1970, Eclipse," Monthly Weather Review, Vol. 97, No. 3, 1970, pp. 280.

41. Thom, H. C. S., and Marcella D. Thom. "Tests of Significance for Temperature and Precipitation Normals," Monthly Weather Review, Vol. 100, No. 6 (June 1972), pp. 503-508.

42. Wisner, Warren M. "The Frequency of Potentially Unfavorable Temperature Conditions in St. Louis, Missouri," NOAA Technical Memorandum NWS CR-53, U. S. Department of Commerce, 1973.

43. Wood, J. L., and W. Alan Bowman. "Cape Kennedy Peak Wind Profile Probabilities for Levels From 10 to 150 Meters," NASA Contractor Report No. 61308, Marshall Space Flight Center, Alabama, 1969.

44. Wood, J. L., B. L. Palmer, B. C. Johnson, and J. E. Tyson. "Extreme Distributions of Ground Wind (3 to 150 Meters) at Cape Kennedy, Florida," Contract NASA8-24927, Marshall Space Flight Center, Alabama, 1970.

45. Yao, Augustine, Y. M. "A Statistical Model for the Surface Relative Humidity," Journal of Applied Meteorology, Vol. 13 (February 1974), pp. 17-21.